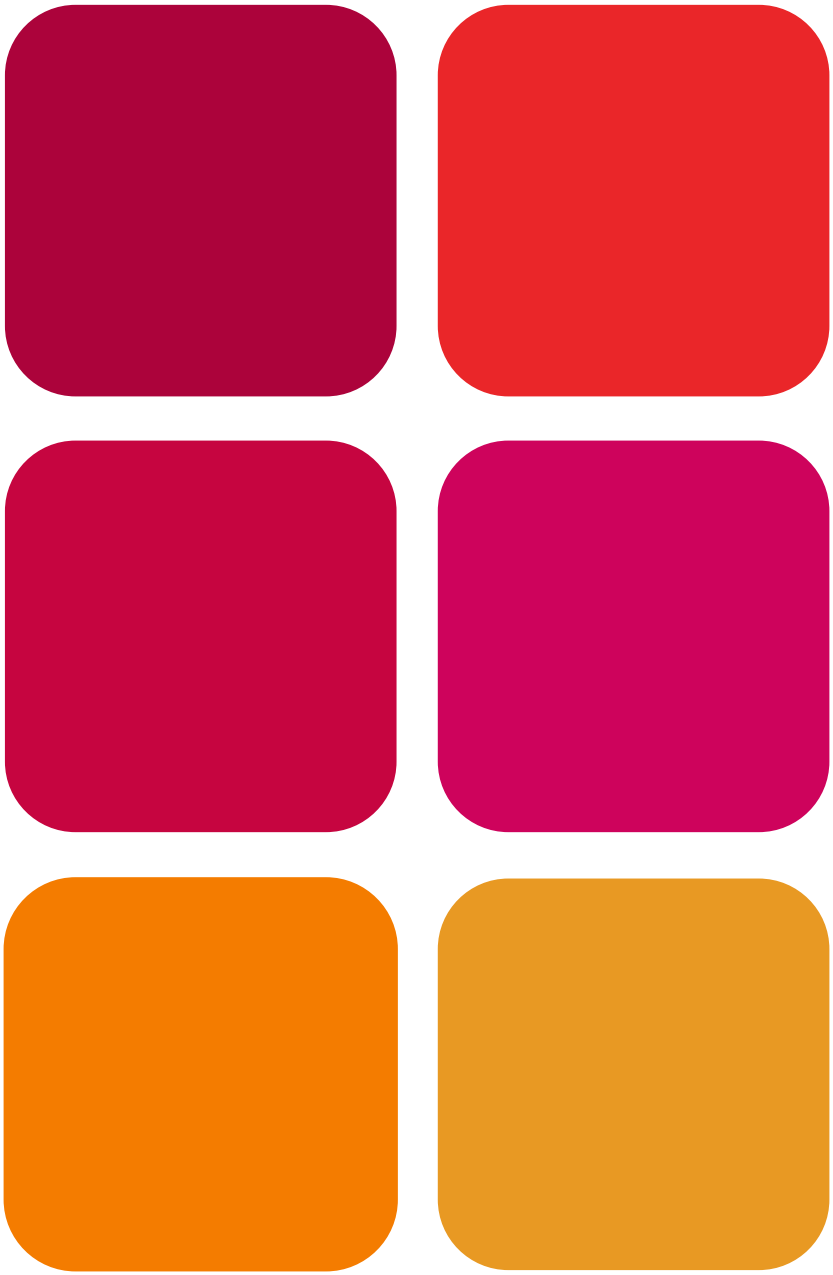


www.sapientia.uji.es | 24

UNIVERSITAT JAUME·I

Introducció a l'estadística aplicada a les ciències socials

Modesto J. Beltrán
María José Peris

Introducció a l'estadística aplicada a les ciències socials

Modesto J. Beltrán
María José Peris

UNIVERSITAT JAUME·I

Departament de Matemàtiques

Codi d'assignatura RA10

Edita: Publicacions de la Universitat Jaume I. Servei de Comunicació i Publicacions
Campus del Riu Sec. Edifici Rectorat i Serveis Centrals. 12071 Castelló de la Plana
http://www.tenda.uji.es e-mail: publicacions@uji.es

Col·lecció Sapientia, 24
www.sapientia.uji.es
Primera edició, 2011

ISBN: 978-84-693-0145-6

Publicacions de la Universitat Jaume I és una editorial membre de l'UNE, cosa que en garanteix la difusió de les obres en els àmbits nacional i internacional. www.une.es

ÍNDEX

Tema 4

Tema 5

Tema 6

Pròleg

L'estudi de l'estadística és, avui en dia, clau en la formació de qualsevol universitari, ja que pot dir-se que el món és ple de gràfics, fórmules, previsions, estimacions, etc., es fonamenten en l'estadística.

El llibre que hem preparat pretén ser, d'una banda, una breu introducció dels continguts més característics d'aquesta branca de les matemàtiques i, d'una altra, un recurs didàctic pràctic que en facilite l'aprenentatge. És per aquest motiu que el text està ple d'exemples detalladament presentats, així com de nombroses explicacions i interpretacions dels aspectes estadístics més complexos. A més a més, s'han obviat, en aquells casos que presenten més dificultat, les demostracions matemàtiques, i se n'han deixat unes quantes de fàcil comprensió.

El públic a qui va adreçat aquest manual és molt ampli, ja que es tracta d'una introducció a l'estadística en la qual no són necessaris grans coneixements matemàtics. És per això que els estudiants dels primers cursos universitaris poden fer ús d'aquest manual. El text està fonamentalment enfocat a estudiants de la branca de les ciències socials, encara que està, evidentment, obert a tothom. Per a acabar, també l'alumnat dels darrers cursos de batxillerat pot fer-ne ús d'alguns capítols.

El contingut del llibre cobreix el procés estadístic elemental. El primer tema és una breu introducció a la investigació social. Es pot considerar un breu resum d'aquest tipus d'estudis i el seu objectiu és contextualitzar el procés estadístic dins de les ciències socials. Es tracta, doncs, d'una enumeració de les característiques principals de tot procés d'investigació social, de les etapes de què consta, de les dificultats principals i de la validesa dels resultats. El segon capítol ubica l'estadística dins d'aquest tipus d'estudis. Així, s'explica en quina part del procés d'una investigació de caire social cal emprar les diferents tècniques estadístiques, les que tracten aquest manual i d'altres.

Els dos temes següents, capítols tres i quatre, conformen el nucli de continguts relatius a l'estudi purament estadístic, per a una variable, des d'un punt de vista descriptiu. És a dir, al llarg d'aquests capítols es comenten els conceptes més elementals de l'estadística descriptiva, com ara taules estadístiques, gràfics, mesures de centralització i mesures de dispersió, entre d'altres.

L'estudi conjunt de dues variables estadístiques es tracta en el capítol cinquè. El tema comença explicant el concepte de *variable bidimensional* i continua revisant aspectes com distribucions marginals, condicionades i correlació lineal. Per a finalitzar el capítol, s'explica el concepte de *recta de regressió* i es mostra com fer-la.

Els temes sis i set formen una revisió de la teoria de la probabilitat clàssica, des del concepte de *atzar* fins al teorema del límit central. En el primer d'aquests dos

capítols es tracten continguts elementals de la probabilitat, com són ara les nocions de probabilitat basada en la freqüència, la probabilitat subjectiva, el teorema de Laplace i les probabilitats condicionades. El tema finalitza amb els teoremes de la probabilitat total i el teorema de Bayes. Pel que fa al segon tema d'aquest grup de dos, s'hi remarquen els aspectes més bàsics dels models de probabilitat. Així, es comenten els conceptes de *variable aleatòria*, i se'n diferencien les discretes i les contínues, i de cadascuna els aspectes més importants. També s'hi tracten els models de probabilitat més característics tant discrets com continus. S'hi posa èmfasi en la distribució normal, per ser aquesta una de les més importants de l'estadística matemàtica.

Els tres capítols següents tracten del darrer aspecte de la teoria de la probabilitat clàssica, la inferència estadística paramètrica. El capítol vuit és una introducció a les distribucions en el mostreig que tenen els estadístics més importants, com ara la mitjana aritmètica o la variància. En el capítol nou es presenten els intervals de confiança. Aquests són intervals construïts a partir de la mostra i permeten assegurar, si es compleixen les hipòtesis, que el paràmetre de la distribució de probabilitat que segueix la població està dins de l'interval amb una confiança determinada. El darrer capítol tracta sobre els contrastos d'hipòtesi. Aquesta tècnica, molt semblant als intervals de confiança esmentats anteriorment, també permet inferir valors per als paràmetres de la distribució de probabilitat de la població.

Finalment, s'hi presenta un conjunt de taules estadístiques necessàries per al càlcul de probabilitats si no és possible consultar software informàtic.

TEMA 1

Breu introducció a la investigació social

Objectius tema 1

- Conèixer els orígens de la investigació social i la seua relació amb els de l'estadística.
- Conèixer els trets més característics d'una investigació social científica.
- Conèixer les dues metodologies fonamentals en la investigació social.
- Conèixer les etapes que cal seguir per a realitzar una investigació social.
- Conèixer el procediment per a mesurar conceptes teòrics.
- Conèixer els trets més importants d'alguns dissenys d'investigació.
- Conèixer els elements que determinen la realització d'una investigació social.

1.1. Introducció

Si es pren la realitat com a font d'informació, s'observa que molt sovint els mitjans de comunicació mostren un munt d'estudis estadístics, molts dels quals fan referència a l'àmbit sociolaboral. Així, els resums d'estudis de mercat, els resultats de les enquestes d'opinió o les reaccions de diferents persones especialitzades respecte a la presentació d'un nou producte apareixen freqüentment tant en la premsa en paper com en l'audiovisual.

Aquests estudis i molts d'altres, que no tenen tant de ressò (molts per ser més locals), són el producte de les decisions preses pels càrrecs directius de les organitzacions. Les raons per les quals es realitzen els estudis són diverses: la detecció d'un problema, la intenció d'una millora de la qualitat d'un procés dins d'una empresa, la necessitat d'innovar, l'avaluació d'un servei, etc.

És, per tant, molt convenient per a qualsevol alt càrrec d'una empresa o organització conèixer o, almenys, tenir unes mínimes nocions de com realitzar una investigació social, de quan s'ha de realitzar, com s'han d'interpretar els resultats, etc. D'altra banda, molts dels estudis de recerca que es realitzen en l'àmbit del treball tenen com a població d'estudi grups socials (treballadors d'una companyia, funcionaris d'una administració, consumidors potencials d'un producte, etc.) o la mateixa societat en conjunt. Consegüentment, les investigacions que es duen a terme són de caire social, amb les peculiaritats i característiques que les distingeixen d'altres tipus de treballs d'investigació.

El procés que se segueix per a elaborar el treball d'investigació es pot esquematitzar de la manera següent: en primer lloc s'ha d'esbrinar el problema que cal investigar, el qual sorgeix per una necessitat descoberta, i establir amb molta claredat els objectius fonamentals. En segon lloc, cal plantejar diverses hipòtesis que solucionen el problema. En tercer lloc, dissenyar la verificació de les hipòtesis plantejades. És en aquest punt quan l'estadística es pot considerar quasi imprescindible. Per a acabar, cal extraure les conclusions de la investigació.

Un exemple es pot trobar a la Universitat Jaume I de Castelló, on l'USE (Unitat de Suport Educatiu) realitza des del curs 94/95 un estudi sobre la qualitat de la docència universitària. L'estudi consisteix a recollir informació, mitjançant una enquesta, sobre diferents trets de la capacitat docent del professorat i sobre el funcionament del procés d'ensenyament en general. La darrera finalitat de la investigació és la millora continuada de l'activitat docent.

Com se'n pot deduir, el procés d'investigació no és trivial, requereix esforç intellectual i experiència per a produir un treball complet d'aquestes característiques, sobretot si s'hi tracten qüestions complexes.

Per a finalitzar aquesta introducció, cal remarcar que aquesta primera unitat no pretén realitzar un estudi acurat i complet del procés d'investigació social –qüestió de

la qual s'ocupa fonamentalment la sociologia–, sinó donar unes nocions bàsiques sobre allò que cal fer si es decideix dur a terme un projecte d'investigació d'aquesta mena. Per tant, aquest contingut no és suficient per a realitzar un treball d'investigació social i caldrà, doncs, complementar-lo amb bibliografia especialitzada.

1.2. Orígens

L'origen comú de l'estadística i les investigacions socials se situa en l'antiga Xina, on l'emperador xinès Xao començava a fer esporàdicament els primers censos de població. També se n'han recollit de les civilitzacions assíria, egípcia i grega. Les finalitats, en aquesta època, d'aquests primitius recomptes de població eren fonamentalment tributàries i militars: els estats necessitaven conèixer els recursos econòmics i militars dels quals disposaven.

En el moviment anomenat *estadística social* que es produí posteriorment, cal ubicar els antecedents més immediats de la investigació social empírica. Concretament als segles XVII i XVIII sorgiren diferents grups de persones que utilitzaren els mateixos procediments emprats en les ciències naturals per a l'estudi de les ciències socials, amb l'objecte de descriure més acuradament la societat de l'època. El vessant més científic d'aquestes investigacions socials fou possible gràcies a les aportacions fonamentals d'una sèrie d'autors englobats en dues escoles estadístiques europees: els aritmètics polítics anglesos i l'escola estadística alemanya. Dels primers, destaquen Graunt, Pretty, Davenant, King i Halley, els quals formaven part d'un corrent pragmàtic constituït per científics naturals que pensaven a quantificar les regularitats socials de la mateixa manera que les naturals. Així, per exemple, Graunt introduí per primer colp les taules d'esperança de vida per a persones de diferents grups d'edats, i hi observà variacions importants entre els habitants d'entorns urbans i rurals; Halley treballà sobre les assegurances de vida; i King va realitzar nombroses aportacions a la demografia moderna. D'altra banda, l'escola estadística alemanya es caracteritzà per la promoció de la denominada *geografia política*, concretament per la comparació de dades demogràfiques, socioeconòmiques i polítiques relatives a diversos països. Entre els màxims representants destaquen Seckendorff, Achenwall i Coring, que va dur a terme més estudis comparatius.

Al llarg del segle XVIII i amb la intenció de mobilitzar l'opinió pública per exigir mesures socials, un conjunt de professionals de diferents camps s'uniren per demostrar empíricament els problemes socials que patien a l'època. Els investigadors constituïren, en els seus respectius països, societats estadístiques, les quals promulgaven el desenvolupament d'enquestes de caire social. Es poden datar, doncs, en aquest període, les primeres investigacions socials empíriques.

Des del començament del segle XIX, els diferents governs anaren creant organismes oficials per a realitzar censos de població: unes enquestes generals que

serviren de registre de la distribució i les condicions de vida de la població en cada país. També es necessitaven conèixer les situacions de l'agricultura i la indústria.

Exemples d'aquestes enquestes socials són:

- *Estadística sobre la moralitat a França* (1833): En aquesta obra s'analitza la relació existent entre la taxa de delinqüència i el nivell cultural.
- *Les condicions morals i físiques de la vida dels obrers de la indústria tèxtil de Manchester* (1832) de James Kay-Schuttleworth. L'autor de l'obra, metge activista de la societat estadística de Manchester que va ocupar el càrrec de secretari del Consell de Sanitat Pública de Manchester, hi descriu les condicions infrahumanes i d'insalubritat en les quals vivien els estrats socials més desafavorits de la població anglesa de principi del segle XIX.

A més d'aquestes enquestes socials, són d'interès per la seua repercussió a l'època, els treballs de Guerry: *L'assaig sobre l'estadística moral de França* (1832) i *L'estadística moral a Anglaterra comparada amb França* (1860). Principalment per com contribuïren al desenvolupament de les estadístiques de la delinqüència.

Aquesta gran evolució de la investigació social al llarg del segle XIX es va produir en part pels avanços en la recollida i l'anàlisi de la informació. Cal destacar-hi Adolf Quételet (1796-1874) i la seua teoria de les regularitats dels fenòmens socials, la qual considera que en la conducta humana hi ha regularitats mesurables mitjançant tècniques estadístiques d'anàlisi. En les investigacions utilitzava també el càlcul de probabilitats. De la mateixa manera, August Comte considerava que en la societat existien trets de base empírica i rigor analític propis de les ciències naturals. Aquesta teoria, l'anomenà *positivisme*, mot que fou molt utilitzat posteriorment. Altres autors importants són Laplace, que introduí el principi dels mínims quadrats en la teoria analítica de les probabilitats; Gauss, que generalitzà el mètode dels mínims quadrats, i Yule, el qual aplicà l'anàlisi de la regressió múltiple i de correlació a les dades socials.

En la segona meitat del segle XIX la recollida de dades socials es va fer més professional i diferents professors universitaris s'interessaren per aquest tipus d'estudis: Gustav Schmoller, Ferdinand Tönies i Max Weber a Alemanya, i Émilie Durkheim a França en són un bon exemple. Aquesta professionalització facilità l'encontre de les dues visions que caracteritzaven els estudis socials empírics: la teoria i l'empirisme. Encara que aquest fet es va produir realment al segle XX.

Els treballs de Durkheim es caracteritzen per concebre la societat com una suma de fets que han d'analitzar-se com a coses. També per considerar que l'estudi de la realitat social ha d'incloure dos tipus d'anàlisi: causal i funcional. Si es vol explicar un fet social, se n'ha de buscar la causa i se n'ha d'indagar la funció. El mètode idoni és el comparatiu: comparar fenòmens en diferents societats. Tanmateix, Max Weber parteix de la distinció entre ciències de la natura i ciències de l'esperit; les primeres se centren en l'observació dels fenòmens naturals, i les altres, en la

interpretació del significat i del sentit de l'acció humana. Aquesta interpretació és possible gràcies a un procés de comprensió: posar-se en el lloc de l'altre. Weber se centra en l'aspecte qualitatiu dels fets socials.

Tot plegat, aquests dos vessants representats per Weber i Durkheim es cristal·litzen en dues visions bàsiques de l'anàlisi de la realitat social: la dimensió estructural i la dimensió intersubjectiva (Rodríguez Ibáñez, 1989). La primera es caracteritza per considerar que la metodologia que cal emprar en les ciències socials és la mateixa que en les ciències naturals, per buscar lleis universals que modelen els fenòmens socials, per fixar-se en els aspectes quantitatius, etc. En canvi, en la dimensió intersubjectiva es diferencien les ciències socials de les naturals per no compartir ni el mètode ni l'objecte de coneixement, per destacar els aspectes qualitatius dels fenòmens socials, per analitzar-ne les qüestions individuals i concretes, etc. Com es dedueix, aquestes dimensions mostren dos models metodològics oposats (quantitatiu i qualitatiu) que amb el transcurs del temps aniran apropant-se.

Al llarg de les primeres dècades del segle XX els investigadors americans i europeus distingien la metodologia qualitativa i la quantitativa per a realitzar els estudis socials. A final dels anys trenta i principi dels quaranta, amb l'arribada als EUA d'alguns membres del Cercle de Viena, confluïren el positivisme lògic d'aquests autors amb el pragmatisme americà. Les crítiques vers la metodologia qualitativa, emprada en l'Escola de Chicago, van suposar el desenvolupament de la metodologia quantitativa. El matemàtic anglès K. Pearson, el francès Poincaré, i en especial, l'instrumentalista John Dewey foren els principals promotors d'aquest impuls.

A mitjan anys quaranta es posen de moda les enquestes, sobretot les d'intenció de vot. Contribuí a aquest fet l'estadístic George N. Gallup, que en 1936 i amb tan sols una mostra de quatre mil cinc-centes persones va donar com a guanyador de les eleccions dels EUA Roosevelt, com així va ser, mentre que la revista *Literary digest* va donar com a guanyador el candidat republicà, Landon amb un sondeig de més de quatre milions d'entrevistes.

Els anys cinquanta i seixanta es caracteritzen per la proliferació d'instituts i fundacions encarregats de crear bases estadístiques, per la introducció dels ordinadors i per la generalització de les investigacions socials, amb un augment de l'especialització. Al començament dels anys seixanta, el funcionalisme perd part de l'hegemonia entre els joves i es produeix el renaixement de les velles teories de Weber. No hi ha cap perspectiva dominant.

La metodologia quantitativa avançà amb l'aparició dels ordinadors. Les crítiques dels qualitativistes defensaren la necessitat de tècniques d'anàlisi diferents de les aplicades a les ciències naturals. Els anys seixanta es van definir com els anys de l'heterodòxia, encarnada per dissidents del paradigma quantitatiu, que havia imperat fins al moment. S'hi produí una pluralitat paradigmàtica.

En el context actual, la investigació social es caracteritza per la síntesi i la integració d'enfocaments macrosociològics i microsociològics.

1.3. La investigació social. Primeres consideracions

L'exemple del llaurador:

Un llaurador té un conreu de taronges que li proporciona alguns beneficis anualment. Un any, observa que la recol·lecció dels fruits no ha anat com esperava, malgrat la climatologia. Neguitós, decideix fer alguna cosa per millorar la collita. Així que pregunta a uns i a altres per què hi ha hagut aquesta disminució de quilograms de taronges i, després de moltes cavil·lacions, creu haver trobat l'explicació a l'esdeveniment: l'excés de taronges que tenien els arbres al començament de setembre n'ha provocat la disminució del nombre de quilograms. Es planteja aleshores la hipòtesi següent: si al començament de setembre hi ha molts fruits per arbre, llavors minva el nombre de quilograms que s'obtenen al final de la campanya.

Decideix fer la comprovació perquè no li torne a passar el mateix. Així, dissenya una estratègia per a comprovar-ho. Escull a l'atzar dos grups del mateix nombre d'arbres i de les mateixes característiques. A un grup, li trau un bon grapat de cítrics al començament de setembre, mentre que, a l'altre, no li'n trau cap. D'aquesta manera pensa: en finalitzar la collita i sempre que tot vaja normalment, podré conèixer si la porga ha fet efecte; tan sols cal calcular els quilograms recollits de tots dos grups d'arbres.

Aquest simple exemple mostra els trets fonamentals d'una investigació. D'una qüestió de la realitat es busquen explicacions, se'n suposa una després de la recerca i, posteriorment, s'intenta demostrar d'alguna manera si les suposicions establertes són certes o no. La investigació científica, però, és molt més complexa i necessita un mètode formal que li done el caràcter científic.

A més a més, és evident que una de les principals qüestions que pretenen totes les investigacions és el rigor i la fiabilitat de les produccions resultants. És per això que, quasi en totes les ciències, el mètode que s'empra per a realitzar les investigacions és el mètode científic, o almenys una adaptació als coneixements que es tracten d'aquest mètode. Cal no oblidar que aquest mètode és el que atorga la categoria de ciència als conjunts de coneixements.

Per aquesta raó es realitzarà un esbós sobre la investigació i el mètode científic.

1.3.1. Investigació i mètode científic

Els diccionaris solen definir el concepte *investigar* com 'tractar d'esbrinar (alguna cosa) indagant-ne i examinant-ne atentament els vestigis, els indicis, etc.'. Per tant, qualsevol diligència dirigida a aclarir o trobar la solució d'alguna qüestió

o problema és una investigació, la qual cosa necessita un mètode d'investigació per a dur-se a terme.

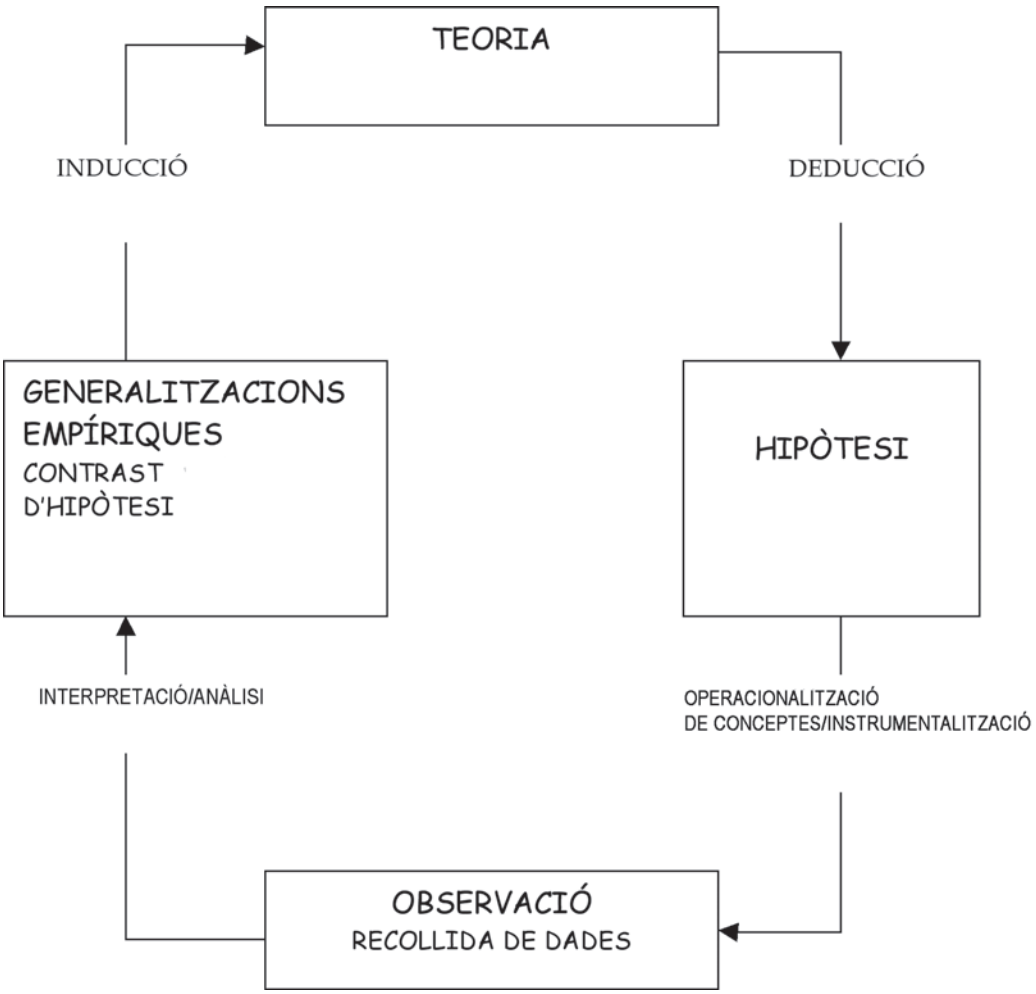
Hi ha diversos mètodes d'investigació (formes d'actuació humana encaminades al coneixement de la realitat observable), únicament han de comprendre un contingut determinat o unes fases d'actuació per a seguir, i una base racional que els sustente, com, per exemple, pressupostos filosòfics del mètode, principis racionals que justifiquen les actuacions que suposa, tècniques específiques per a dur-ne a efecte les etapes, etc.

De tots aquests mètodes, el mètode d'investigació per excel·lència és el mètode d'investigació científic, ja que el seu objecte de coneixement és la realitat, comprèn les premisses de tot mètode d'investigació i disposa de tècniques apropiades. A més, atès que té un caràcter instrumental per ser mètode, pot ser emprat per totes les ciències: les fisiconaturals que l'originaren, les socials i les humanes.

Pel que fa a les ciències socials, cal remarcar que un dels esquemes més coneguts en la comunitat científica que es dedica a les investigacions d'aquest tipus és el que Wallace publicà en 1971, el qual descriu un procés circular en què es condicionen mútuament la teoria i la investigació (quadre 1). El procés comença en la teoria, de la qual es trauen, mitjançant deduccions lògiques, unes hipòtesis concretes que cal operacionalitzar per a ser contrastades empíricament. Després, es procedeix a l'observació o a la recollida de dades, les quals són sotmeses a anàlisi. Finalment, la informació extreta de la investigació s'interpreta. D'aquesta interpretació es dedueixen unes generalitzacions empíriques que són contrastades amb les hipòtesis plantejades.

Si les dades empíriques contradiuen les hipòtesis de l'estudi, la teoria de la qual es van deduir es refusa. En cas contrari, s'accepten les hipòtesis. Aquesta acceptació no n'implica una verificació teòrica sinó una confirmació.

D'aquesta manera, les generalitzacions empíriques reverteixen, mitjançant el mètode inductiu, en la teoria, que es veu refusada, acceptada o lleugerament modificada. Si la teoria no és recolzada per les dades empíriques, el procés comença de nou; per tant, en el procés d'adquisició del coneixement científic es contempla un *feedback* continuat.



Quadre 1

1.3.2. La investigació social científica

La ciència es pot definir com un conjunt de coneixements sobre la realitat observable obtinguts pel mètode científic. Les ciències socials tenen com a objecte d'estudi la realitat social. A més a més, és evident que en l'estudi de la societat es pot aplicar el mètode científic: s'hi poden formular qüestions o problemes referits als fenòmens socials, intuir-ne solucions i verificar-les.

La investigació social científica consisteix, precisament, a realitzar recerques sobre la societat emprant el mètode científic. Així, es pot definir com el procés d'aplicació del mètode i les tècniques científics a situacions i problemes concrets en l'àrea de la realitat social per a buscar-ne respostes i obtenir nous coneixements (Sierra, 1995).

D'altra banda, és evident que no sempre és senzill aplicar el mètode científic en aquesta ciència per les particulars característiques dels fenòmens socials: són en moltes ocasions qualitatius, estan influïts per múltiples factors, són molt variables en el temps (la qual cosa dificulta la recerca de regularitats) i manquen d'instruments potents d'observació. Aquestes qüestions i d'altres referides als valors i les creences

socials i de l'investigador, són algunes de les dificultats que presenta la investigació científica de la societat.

Malgrat aquestes dificultats, és el mètode científic, i no alguns dels alternatius, com el dialèctic i el funcionalista, el que s'usa fonamentalment. No obstant això, el mètode científic no s'utilitza de la mateixa manera en totes les investigacions. Així, no s'analitzaran de la mateixa manera un estudi sobre la intenció de vot que una investigació sobre els sentiments afectius d'una mare amb el fill, ni s'hi empraran les mateixes tècniques.

Això fa que existisquen diferents modalitats d'aplicació del mètode científic segons les característiques de l'objecte d'estudi. El mètode s'ha d'adequar a l'objecte. Així mateix, tot i la pluralitat existent, s'hi distingeix una dicotomia metodològica que es basa en dues perspectives sobre el mètode que cal emprar en la investigació social:

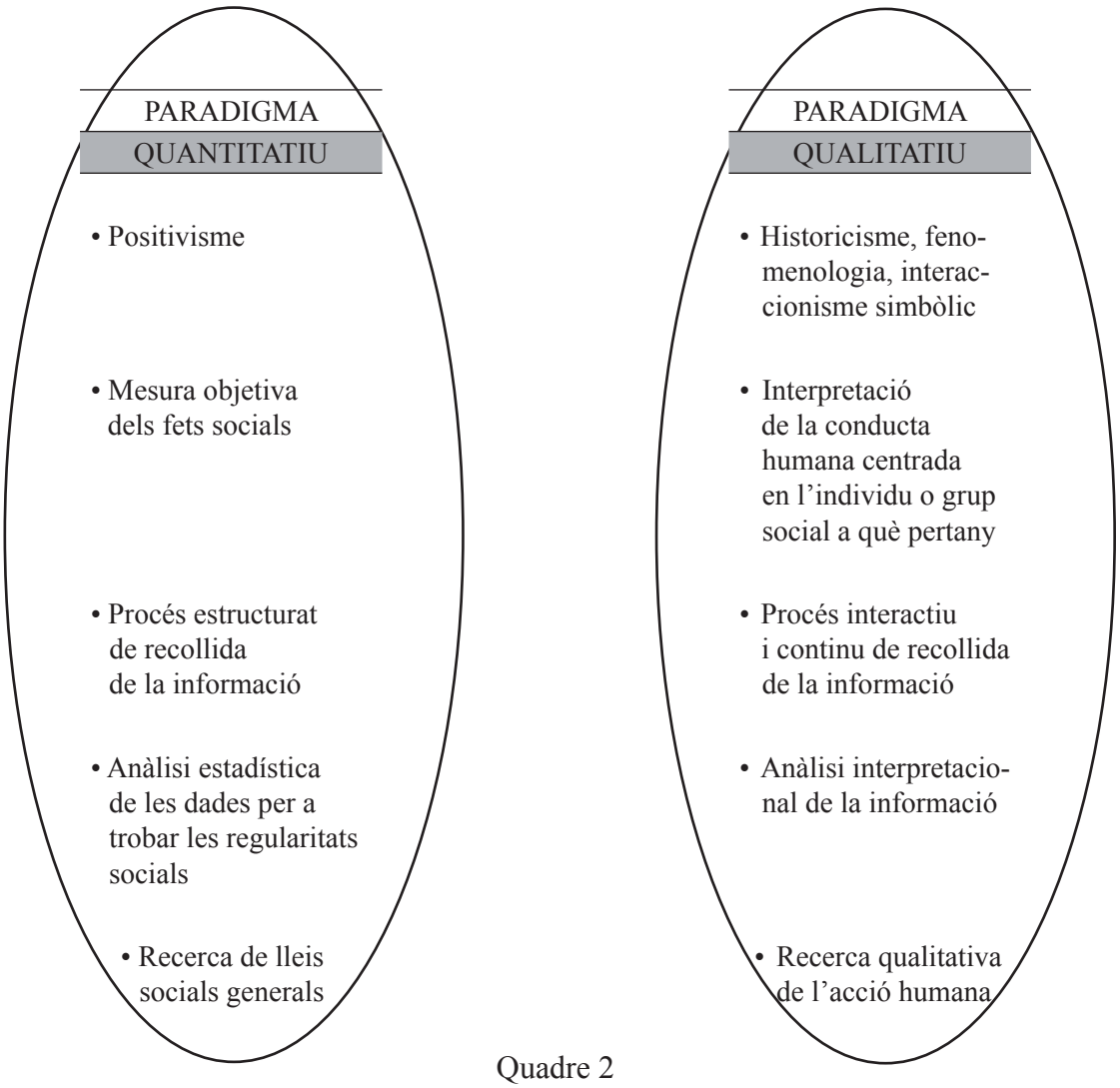
- La perspectiva humanista/qualitativa: concep l'especificitat de les ciències socials. Rebutja el mètode emprat en les ciències naturals i advoca per l'anàlisi d'allò individual i concret, mitjançant la comprensió o la interpretació dels significats intersubjectius de l'acció social (des del punt de vista de l'actor social). L'èmfasi es posa en el llenguatge i els aspectes micro de la vida social (situacions cara a cara).

- La perspectiva científica/quantitativa: defèn l'existència d'un únic mètode (el de les ciències naturals i exactes) general de totes les ciències, al mateix temps que el principi de causalitat (explicació dels fenòmens socials mitjançant les causes que els produeixen) i la formulació de lleis generals en l'anàlisi de la realitat social. L'èmfasi es posa en l'explicació, en la contrastació empírica i en el mesurament dels fenòmens socials.

Encara que diversos autors reconeixen alguns mètodes més d'estudi dels fenòmens socials (el mètode històric, el comparatiu i el critico-racional), el cert és que els dos mètodes mencionats amb anterioritat són els fonamentals. Tots dos es diferencien tant en l'estratègia seguida en el reconeixement de la informació com en la seua anàlisi, a causa de les diferents perspectives paradigmàtiques, les quals proporcionen un marc filosòfic i metodològic concret per a estudiar la realitat social (Filstead, 1986).

Un paradigma representa un model fonamental, «una imatge bàsica de l'objecte d'una ciència. Serveix per a definir el que s'ha d'estudiar, les preguntes que és necessari respondre, com s'han de preguntar i quines regles s'han de seguir per interpretar les respostes obtingudes» (Ritzer, 1993).

El quadre següent resumeix les característiques diferenciadores que tradicionalment s'han atribuït als dos paradigmes en qüestió. Cal dir, però, que els atributs resumits tenen un caràcter genèric i no estan exempts de controvèrsia.



Quadre 2

Com es dedueix, les diferències bàsiques entre les dues modalitats d'aplicació del mètode científic són: d'una banda, el caràcter numèric del mètode quantitatiu en contrast amb la informació en llenguatge natural del qualitatiu al llarg de les tres etapes del mètode científic (Schwartz, 1984); i, d'altra banda, que les interpretacions i les explicacions en el mètode quantitatiu són més objectives que en el qualitatiu, ja que en aquest darrer es fonamenten en la comprensió íntima de la realitat per part de l'investigador.

Pel que fa a alguns exemples, els estudis de mercat solen emprar tècniques quantitatives, ja que tenen la necessitat de conèixer i quantificar aquelles variables del mercat que són de vital importància per a l'empresa-producte: coneixement de les marques, hàbits de compra i consum, motius del consum, etc. Les tècniques de recollida de dades solen ser enquestes minuciosament preparades i l'anàlisi de les dades obtingudes sol ser estadística. Els resultats facilitaran a l'empresa informació sobre les característiques del mercat del producte.

D'altra banda, si una empresa decideix investigar sobre nous conceptes de productes, nous camins d'investigació o si vol observar i investigar comportaments, conductes, motivacions, actituds... sobre una marca o un producte, pot escollir una tècnica qualitativa com les de les dinàmiques de grup (Philips 66, Brainstorming,

etc.). Així, una vegada definits els objectius escollits, la grandària, els components del grup i les tècniques que s'utilitzaran, es realitzen les reunions. Posteriorment se n'analitzen els continguts (normalment gravacions) amb l'objecte d'establir-ne les conclusions. La informació ha de ser interpretada pels especialistes en dinàmiques de grup, òbviament.

Per a finalitzar el punt es mostrarà un exemple en què es pot realitzar un estudi quantitatiu o un de qualitatiu: el test de l'envàs. Pot exemplificar-se en cas que una empresa desitge analitzar l'envàs d'un producte per a lactants. Es podria optar per utilitzar tècniques grupals amb mares recents, o per realitzar una enquesta a un sector delimitat de la població. Del primer estudi, probablement, s'obtindrien resultats que, d'una banda, permetrien innovar l'envàs i, de l'altra banda, no serien generalitzables a la resta de la població. Tanmateix, en el segon cas els resultats sí que serien generalitzables però no originarien tantes innovacions. Cal dir que, en algunes ocasions, en primer lloc es realitza un estudi qualitatiu i després, un de quantitatiu per contrastar els resultats de l'anterior.

Cal remarcar que aquest text se centra sobretot en els estudis quantitatius, per estar més relacionats amb l'estadística que no pas els qualitatius.

1.4. El procés seguit en una investigació social

En els punts anteriors s'ha posat de manifest que la investigació social és un procés que pretén conceptualitzar la realitat objecte d'estudi, és a dir, obtenir-ne coneixements, idees i representacions intel·lectuals tan exactes com siga possible. És mitjançant un seguit d'actuacions successives i interrelacionades com es poden aconseguir els objectius cercats en la investigació.

En aquest procés global es poden distingir, seguint Sierra Bravo (1995), dos processos. El procés metodològic determina les etapes que cal prosseguir en la recerca de la solució al problema plantejat. El procés lògic segueix els principis del mètode científic i es fonamenta en els elements conceptuals que intervenen en la investigació social, així com en les seues interrelacions.

1.4.1. Etapes

Cal entendre, en aquest apartat, el mètode, com el conjunt d'etapes que cal seguir per a arribar a l'obtenció dels resultats. Des d'aquest punt de vista i seguint Bunge (1969), les etapes del procés metodològic en un projecte d'investigació social són:

- Descobrir el problema que s'investigarà.
- Documentar i definir el problema.
- Imaginar-ne una resposta probable o hipòtesi.
- Deduir o imaginar conseqüències de les hipòtesis o subhipòtesis empíriques.
- Dissenyar la verificació de les hipòtesis o del procediment concret que s'han de seguir en la prova.
- Posar a prova o contrastar amb la realitat les hipòtesis a través de les seues conseqüències o mitjançant subhipòtesis empíriques.
- Establir les conclusions resultants de la investigació.
- Estendre les conclusions o generalitzar els resultats.

El problema, origen de la investigació, consisteix en una pregunta o un interrogant sobre la realitat. Amb la investigació social es busca la solució. Es tracta d'una activitat complexa que suposa la resposta a una pregunta: què es desitja saber? La resposta inicial a aquesta pregunta és normalment vaga i difusa, i és necessari precisar-la en les etapes successives.

La concreció del problema s'obté com a conseqüència d'un estudi a fons del tema que es tracte per a tenir a l'abast la màxima quantitat d'informació sobre allò que es pretén investigar. D'aquesta manera, es pot enunciar el problema més concretament i, al mateix temps, fixar els objectius de la investigació.

A continuació s'imaginen les solucions més probables o hipòtesis. Les hipòtesis determinen l'objecte de verificació i n'orienten totes les fases. La contrastació de les hipòtesis es realitza normalment, no d'una manera directa, sinó mitjançant la imaginació i la deducció lògica de conseqüències empíriques concretes, que, si són immediatament verificables, s'anomenen *subhipòtesis*.

El disseny de la investigació ha d'especificar i planificar la manera concreta de verificar les hipòtesis. Cal que indique com s'ha de provar que la hipòtesi és vertadera i quina pauta cal seguir en la recollida i el tractament de les dades. Entre les operacions concretes d'aquesta fase destaquen:

- L'especificació de les variables i les relacions.
- La determinació d'altres variables no estudiades, que poden influir en les resultats i preveure els procediments per al seu control.
- La concreció de les dades necessàries sobre les variables investigades (determinació d'on s'obtindran, com recollir-les i com tractar-les).

Abans d'iniciar el treball de camp, cal efectuar l'elecció de les tècniques d'observació i la construcció d'instruments per a fer la recollida de dades. La forma de tractament de les dades demana la previsió de les taules de dades necessàries i els tipus d'anàlisi, estadística o d'altres. Les operacions bàsiques de la prova són l'observació, la classificació i l'anàlisi.

Una vegada obtingudes les conclusions, es comparen amb les hipòtesis formulades i la teoria que fonamenta la qüestió investigada. A continuació segueix el procés expositiu dels resultats, que es concreta en l'informe, en el qual s'exposaran el mètode, el procés d'investigació i els resultats.

Per a finalitzar, cal remarcar que no totes les investigacions empíriques de caire social segueixen les etapes esmentades anteriorment i de forma estrictament ordenada, depèn de l'estudi i de la metodologia seguida (qualitativa o quantitativa). Així, en les investigacions qualitatives es produeix una redefinició constant de l'objecte d'estudi i de les hipòtesis fins que es troben les adients. En les quantitatives no és tan habitual aquest aspecte però tampoc se segueixen sempre i de forma rígida les fases que hem assenyalat adés.

1.4.2. Lògica

El procés lògic es fonamenta en el mètode científic i és paral·lel al metodològic. S'hi tenen en compte els elements conceptuals que intervenen en la investigació i les seues relacions.

En aquest procés, i de manera semblant al mètode científic, es poden distingir dos subprocessos de moviment invers: el de verificació i el de teorització. El primer comença en les idees i finalitza en la realitat; per contra, el segon parteix de la realitat i finalitza en les idees.

El procés de verificació

Com el seu nom indica, el procés de verificació és una provatura en la realitat d'allò teòric. És un procés principalment deductiu. Els elements que el conformen són: la teoria i els models, les hipòtesis, els fets i la verificació.

La teoria es pot definir com un conjunt de proposicions connectades lògicament i ordenada que intenten explicar una zona de la realitat mitjançant la formulació de les lleis que la regeixen (*Diccionario de las Ciencias Sociales*, 1976). El seu estudi durant la investigació es realitza després que s'haja establert allò que es desitja investigar. D'altra banda, les proposicions que formen els resultats teòrics faciliten l'obtenció dels indicadors o les variables necessaris per a descriure l'objecte d'investigació, així com de les relacions existents entre les mateixes variables. Per a representar i formalitzar les relacions s'empren els models, que són proposicions aplicables a moltes teories i que permeten escriure matemàticament les relacions de les variables i la concreció de la teoria.

Aquest costós treball de recerca facilita a l'investigador concretar el problema que cal investigar i, en conseqüència, li permet generar un seguit de possibles solucions o hipòtesis, les quals poden ser alternatives o no. Les hipòtesis són, doncs, enunciats que es caracteritzen per ser idees suposades no verificades però probables, i per referir-se a variables o relacions entre les variables. Per a representar-les s'usen els models: els gràfics (s'hi relacionen les variables mitjançant fletxes), els matemàtics (especificació matemàtica de les relacions entre les variables), etc.

Pel que fa als fets, es poden definir com tot allò que existeix en la realitat, independentment del pensament humà.

La verificació és l'element central del procés. Consisteix en un conjunt d'actuacions que relacionen les hipòtesis i els fets. Deixant de banda les discrepàncies sobre si la verificació determina la demostració total de les hipòtesis o la no-oposició a la realitat, el cert és que la prova pot tenir dos resultats possibles. Si és positiu vol dir que les hipòtesis han quedat confirmades pels fets de la realitat investigada. Tanmateix, si és negatiu cal refutar les hipòtesis o bé reformular-les.

El procés de teorització

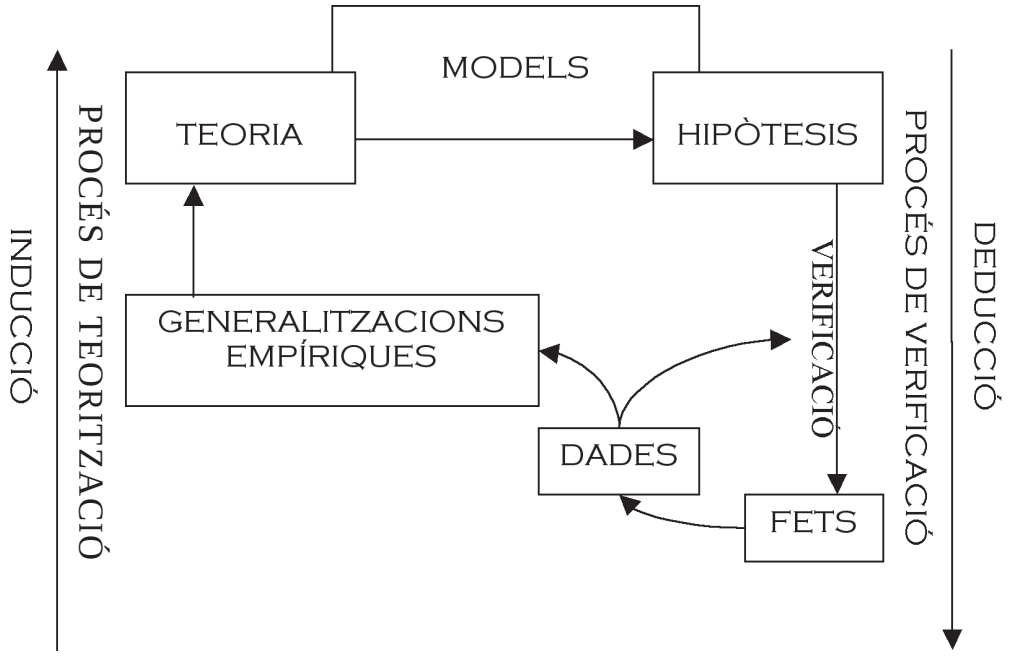
El procés de teorització, més inductiu, comença en els fets i recorre el camí invers, és a dir, dels fets fins a la teoria. Per tant, tots dos processos segueixen camins oposats. Els elements que el formen són les dades, la formació de proposicions i la teoria.

Les *dades* són els materials més simples de la investigació científica. No són fets, sinó expressions d'aquests. A més, les dades exclouen tota inferència o deducció, són el resultat mecànic de l'observació.

Analitzant les dades, agrupant-les i relacionant les variables de les quals provenen, s'obtenen generalitzacions empíriques. Per altra part, aquesta anàlisi també pot confirmar les hipòtesis, que en aquest cas passen a ser enunciats científics.

Si les generalitzacions empíriques apareixen confirmades per moltes investigacions i representen regularitats vàlides per a tota la població en general, reben el nom de *lleis científiques* i formen part de la teoria.

El quadre següent resumeix el que hem esmentat:



Quadre 3

1.5. Qüestions bàsiques en un projecte d'investigació social

Fins al moment, s'han tractat els elements principals que intervenen en un procés d'investigació social, així com les etapes que aquest ha de complir des d'un punt de vista molt teòric, fixant-se únicament en qüestions purament conceptuals. Cal tenir present, però, que no són les úniques que hi intervenen, també aspectes econòmics i temporals determinen el procés. Així, en la pràctica habitual, el projecte d'investigació té en compte tres elements claus:

- Els objectius de la investigació.
- Els recursos o mitjans materials, econòmics i humans dels quals disposa l'investigador per a realitzar l'estudi.
- El temps concedit per a realitzar-lo.

La conjunció d'aquests elements marcarà el disseny de la investigació (pla global de la investigació que integra, de manera coherent i adequada, tècniques de recollida de dades, anàlisis previstes...) (Alvira, 1989). Cal diferenciar aquest concepte del de projecte en l'àmbit de la investigació social, ja que aquest últim té una connotació molt més àmplia.

És necessari assenyalar que el procés d'investigació no ha de concebre's com un procediment fix. Malgrat que el projecte determine un conjunt d'actuacions per al compliment de cada fase d'indagació, la posada en pràctica pot dur a modificacions.

A continuació, i seguint Cea d'Ancona (1996), s'esmenten resumidament els elements essencials d'un projecte d'investigació social. A alguns d'aquests s'ha fet ja referència en els epígrafs anteriors.

1.5.1. La formulació del problema d'investigació

Bàsicament, en aquest primer punt cal definir i emmarcar allò que es desitja investigar. En primer lloc, és necessari establir exactament el que es pretén analitzar: quins són els objectius generals i específics de la investigació. Aquests objectius provenen d'una idea original, donada o descoberta, la qual, en un principi, no és més que això, una idea.

Amb la finalitat de precisar-la i configurar-la, caldrà elaborar el marc teòric de la investigació, centrar els objectius que han sigut determinats per la primera idea en la literatura que els tracta. Caldrà, doncs, realitzar una revisió bibliogràfica profunda. D'aquesta manera, serà possible analitzar i discernir si la teoria existent i la investigació prèvia suggereixen alguna resposta a les preguntes de la investigació o una direcció a seguir dins l'estudi. Al mateix temps, sembla coherent

realitzar una altra sèrie d'investigacions exploratòries: discutir el tema amb altres investigadors, a fi que puguen aportar idees; entrevistar alguns dels implicats en el problema que cal investigar, amb el propòsit de recollir informació que ajude en el disseny, etc.

Aquesta fase prèvia és molt important dins del projecte. Del rigor amb què es realitze aquesta fase en dependrà el bon desenvolupament.

En l'exemple que es refereix a la docència a l'UJI, esmentat en la introducció, es pot considerar que la finalitat és la millora de l'activitat docent. Per aconseguir-la se segueixen quatre objectius principals:

- Reconèixer la competència docent del professorat.
- Identificar les necessitats del professorat.
- Col·laborar amb el professorat per superar les necessitats detectades.
- Detectar altres problemàtiques que, potser, influeixen en el desenvolupament normal de la docència.

La pregunta que origina aquesta investigació és la següent: és possible una millora en la competència docent del professorat de la universitat? La resposta a aquesta pregunta origina la hipòtesi de la investigació: sí que és possible aquesta millora. Les subhipòtesis són: el coneixement del professorat dels resultats d'una avaluació de la seua tasca docent provocarà una millora en l'activitat docent; la identificació de les necessitats del professorat junt a la col·laboració de l'USE tindrà com a conseqüència la millora de la docència per part del professorat.

En aquesta tasca la investigació té caire descriptiu, ja que fonamentalment es tracta de conèixer com realitza l'activitat docent el professorat de l'UJI. El propòsit final és, clarament, provocar l'autoreflexió en el professorat per tal que, amb l'ajuda de l'USE, puga millorar la seua activitat a la universitat.

1.5.2. Unitats d'observació

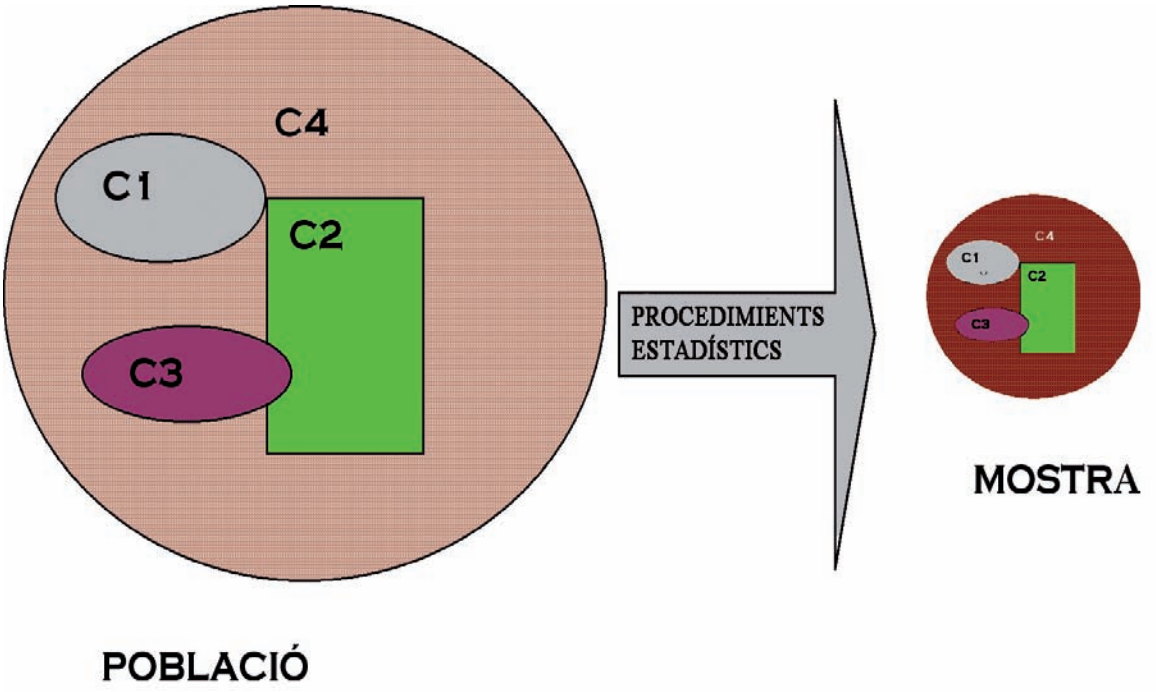
La població d'estudi o unitats d'observació són les realitats de les quals es pretén estudiar alguna cosa. Són claus en la investigació, ja que és d'ací d'on s'obtenen les dades empíriques necessàries per a contrastar les hipòtesis amb la realitat. Cal recordar que els elements estructurals de les hipòtesis són les variables, la població d'estudi i les relacions que les uneixen entre si.

D'altra banda, és convenient tenir present que en poques ocasions s'estudien totes les unitats que formen la població. Per raons econòmiques i temporals, per a realitzar l'estudi, se solen escollir uns quants elements de la població –tan representatius com siga possible. El conjunt format per totes aquestes unitats s'anomena *mostra* i la seua elecció és crucial perquè els resultats de la investigació siguen vàlids.

L'ús d'una mostra presenta notables avantatges; no obstant això, cal tenir present que la mostra ha de complir una sèrie de condicions. La grandària ha de ser

estadísticament proporcional a la de la població, l'elecció dels elements que la formen no ha d'estar esbiaixat, si no es vol que la mostra presente aquesta característica, etc. A més a més, la mostra ha de ser un fidel reflex de l'univers que representa, és a dir, tots els grups en què es puga dividir la població han de ser a la mostra. En certa manera, la mostra ha de ser la població a petita escala (quadre 4). Existeix un conjunt de procediments estadístics que faciliten el control sobre aquestes condicions, i que seran tractats en unitats posteriors.

Tal com destaca Galtung, l'elecció de les unitats d'observació és probablement la primera elecció decisiva en qualsevol investigació. Un colp realitzada, és molt complex tornar enrere perquè tot el procés d'investigació s'haurà edificat sobre aquesta elecció. Es pot dir el mateix de l'elecció de la mostra.



Quadre 4

1.5.3. L'operacionalització del problema

Un colp definit i delimitat l'objecte d'estudi, cal concretar-lo. És necessari establir els conceptes, les categories i les variables que s'analitzaran.

A més a més, els conceptes s'han de transformar en variables o indicadors (observables o que es manifesten) que en possibiliten la contrastació empírica. Aquest procés es realitza amb la finalitat de mesurar el que s'està investigant, és a dir, formalitzar matemàticament les propietats latents emmarcades en el concepte.

També les hipòtesis, que com ja s'ha esmentat amb anterioritat avancen respostes probables a les preguntes inicials de la investigació i s'expressen en forma de proposicions, han d'expressar-se en termes d'indicadors, ja que mostren relacions entre conceptes, valors hipotètics o estimats de variables, etc.

La dificultat de mesurar

En quasi totes les investigacions socials és necessari mesurar, al llarg del procés, una sèrie de qüestions. Si aquestes són magnituds o variables que tenen associades un procediment evident de mesura, com poden ser l'edat, el pes o el sou mensual, la tasca és summament senzilla. El problema sorgeix quan es volen mesurar conceptes que no tenen aquestes característiques, com per exemple les relacions interpersonals existents entre els empleats d'una empresa o el nivell cultural d'una població. En aquests darrers casos cal operacionalitzar els conceptes.

El terme *operacionalitzar* és l'usat per a representar el procés d'assignacions de mesures a conceptes. En certa manera consisteix a transformar els conceptes en variables que es puguen mesurar, entenent *mesurar* –i seguint Carmines i Zeller– com el procés de vincular indicadors empírics als conceptes abstractes.

Els conceptes no són directament mesurables (no són observables), sinó que es troben representats per respostes (a preguntes referents al concepte) que sí que ho són. Per exemple, «nombre d'amics que té un alumne a l'escola» es pot considerar com un indicador del concepte *l'amistat a l'escola*. D'altra banda, convé remarcar que els conceptes, malgrat les seues diferències respecte al grau d'abstracció («amor» és més abstracte que «educació»), sintetitzen diferents aspectes observables que conjuntament els defineixen: menjar poc, tenir mal de cap, esternudar, tenir son, parlar poc, etc., són manifestacions que determinen el concepte *malaltia*, per exemple. Aquestes manifestacions faciliten l'obtenció dels indicadors.

Així doncs, la concreció del concepte necessita ser traduïda a indicadors, variables empíriques (observables o manifestes), que mesuren les propietats latents del concepte.

Les variables

Per *variable* se sol entendre la qualitat o característica observable d'un objecte o esdeveniment que té almenys dues categories o valors diferents. Aquests atributs classifiquen l'objecte o l'esdeveniment i en permeten la mesura. A més a més, si no es consideren aïlladament, les variables es poden relacionar entre si, de manera que la modificació d'una pot influir en el valor que prenen unes altres.

De la primera noció es poden donar els exemples de la variable «pes» (kg), que classifica les persones segons els quilograms de massa que tenen; la variable «nombre de fills» (nombres naturals), que classifica les persones segons el nombre de descendents; la variable «partit polític afí» (diferents formacions polítiques d'un país), segons l'afinitat política, etc. Respecte a la segona noció, les variables són característiques observables d'alguna cosa, lligades entre si per una relació determinada, com ara la covariància o associació, la dependència, la influència o la causalitat. Per exemple, la variable «nivell cultural d'una persona» està lligada, segons una relació de dependència, amb altres variables com poden ser l'edat, el nivell social, la intel·ligència, els ingressos, etc.

D'altra banda, les variables es poden classificar segons diferents modes: nivell de mesurament, escala de mesurament, funció en la investigació i nivell d'abstracció. Com que les darreres classificacions tenen relació amb el punt següent, es tractaran seguidament, i es deixarà per a la unitat 3 la resta de la tipologia de les variables.

Així doncs, segons el paper que tenen en la investigació, es poden classificar en:

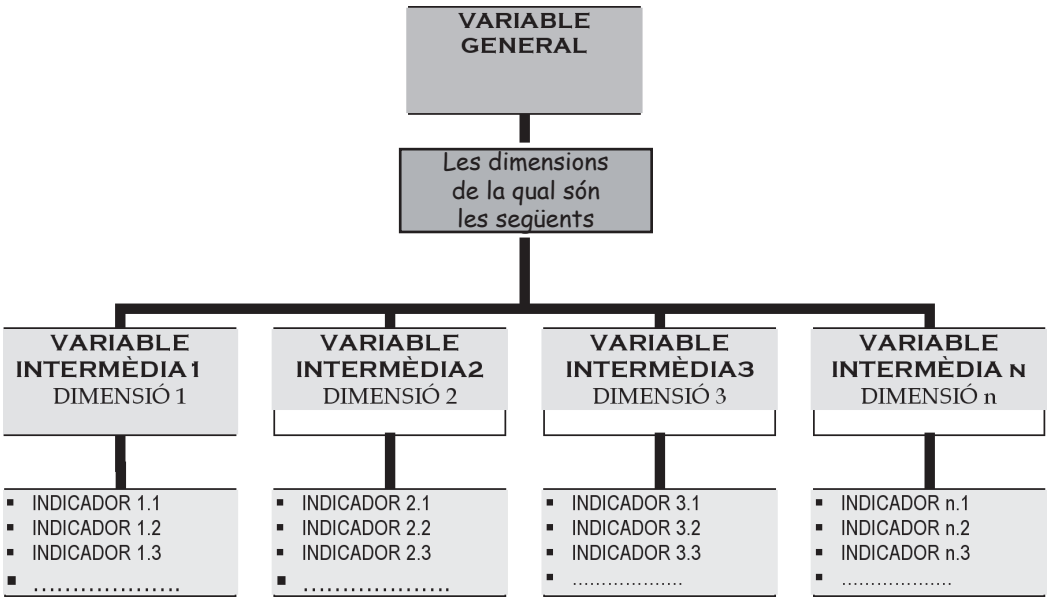
- Variables dependents: són aquelles que es pretenen explicar. Per a fer-ho, s'estudia la influència que sobre els seus resultats tenen unes altres variables, les variables independents.
- Variables independents: són les variables explicatives, que se suposa que influeixen en els resultats de les variables dependents. Aquesta influència o associació amb les variables dependents centra l'interès de la investigació.
- Variables externes o estranyes: són aquelles variables que poden influir en la relació existent entre les variables dependents i les independents. Aquestes variables poden produir explicacions alternatives dels valors que prenen les variables dependents (diferents de les donades per les variables independents). És, per tant, necessari controlar-ne els efectes, bé abans de realitzar la recollida de dades o bé després.

Per exemple: es desitja conèixer la relació existent entre les hores d'estudi i la nota obtinguda en una oposició.

En aquest simple exemple, la nota obtinguda constitueix la variable dependent i les hores d'estudi, la independent. D'altra banda, si de la informació recollida es conclou que un alt percentatge de persones que tingueren una nota alta en l'oposició havien estudiat moltes hores, per a afirmar que la nota obtinguda depèn de les hores d'estudi, cal haver controlat unes altres variables que poden influir en aquesta relació. Per això serà necessari indagar en variables que incidisquen diferencialment en els opositors amb similars puntuacions obtingudes en l'oposició. Així, variables com l'edat dels opositors, el nivell intel·lectual, el nombre de vegades que s'hi presenten, etc., caldrà analitzar-les també.

I segons el nivell d'abstracció (quadre 5), les variables es poden classificar en:

- Variables generals: són aquelles que pel nivell d'abstracció no poden ser directament observades. Per a mesurar aquestes variables cal traduir-les a variables intermèdies i indicadors. En pot ser un exemple el nivell econòmic d'un país.
- Variables intermèdies: expressen aspectes o dimensions parcials de les generals i, per tant, més properes a la realitat. En l'exemple anterior, el desenvolupament industrial.
- Variables empíriques o indicadors: representen aspectes específics de les variables intermèdies (dimensions), que es poden mesurar directament. En l'exemple anterior, el consum d'energia elèctrica per habitant.



Quadre 5

Els conceptes teòrics

Un concepte teòric es pot considerar com una variable general i, per tant, no és directament observable. És per això que cal establir una definició operativa del concepte una vegada se'n coneixen clarament els principals trets definitoris. És a dir, cal concretar les abstraccions del concepte mitjançant variables empíriques que el permeten mesurar.

El procediment de transformació de variable general a empírica fou desenvolupat per Paul F. Lazarsfeld, que en distingia les fases següents:

1. Representació teòrica del concepte de manera que quede reflectit en una noció teòrica els trets principals que representen en realitat.
2. Especificació del concepte, descomponent-lo en diferents dimensions o en els aspectes més rellevants que engloba. Per exemple, Lazarsfeld assenyala que en la noció de rendiment es poden distingir tres dimensions: ritme de treball, qualitat del producte i rendibilitat de l'equip.
3. Per a cada dimensió, selecció d'una sèrie d'indicadors que mostren l'extensió que aconsegueix la dimensió en els casos investigats. Per exemple, una variable intermèdia o dimensió de la variable general «classe social», és el nivell econòmic, i indicadors d'aquest són l'import de totes les fonts de renda: sous, rendiments de finques, interessos de capitals, deutes...
4. Construcció d'índexs que sintetitzen els indicadors. Com que cada indicador no té la mateixa importància dins de la investigació, en aquesta fase se li assigna un pes o valor. A partir d'aquests valors es confecciona un índex, una mesura comuna que agrupe diversos indicadors d'una mateixa dimensió conceptual.

Els indicadors constitueixen propietats essencialment manifestes que es troben empíricament relacionades amb la propietat latent no observable (dimensió). És per això que se'ls exigeix que siguen «expressió numèrica, quantitativa de la dimensió que reflecteixen» (González Blasco, 1989).

Com deia Durkheim, «per a estudiar la solidaritat, un fet intern, cal substituir-la per indicadors que la simbolitzen».

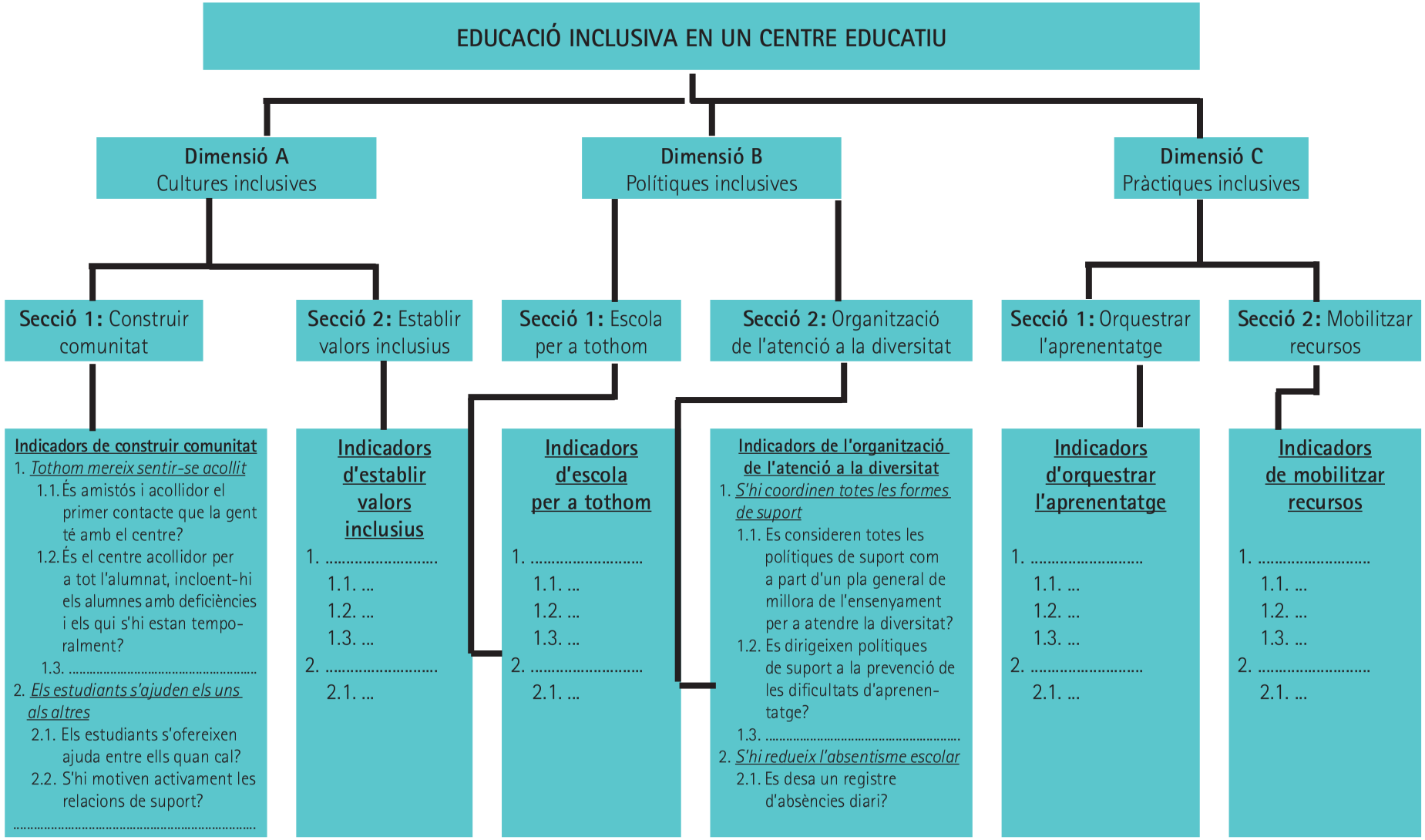
Per a finalitzar aquest apartat relatiu als indicadors, s'ha de remarcar, en primer lloc, que aquests estan donats en termes de probabilitats, ja que representen aproximacions als conceptes que mesuren. En segon lloc, una vegada decidits, cal comprovar fins a quin punt són vàlids i fiables.

La validesa fa referència a l'adequació dels indicadors escollits per a representar els conceptes teòrics, és a dir, l'investigador ha de comprovar que els indicadors mesuren correctament el significat donat al concepte teòric en consideració. D'altra banda, la fiabilitat fa referència a la consistència dels resultats en el temps. És a dir, els resultats obtinguts en successius mesuraments del concepte teòric han de ser iguals per a asseverar-ne la fiabilitat.

Exemple 1: De la variable general als indicadors: l'educació inclusiva

Darrerament està en voga l'educació inclusiva, la qual implica un canvi en els principis que sustenten l'estès sistema educatiu selectiu. Així, l'objectiu fonamental de l'educació inclusiva és formar, a més a més d'una persona hàbil i competent, una persona autònoma i crítica que sàpia com aprendre i conviure en societat, i que desenvolupe al màxim totes les seues capacitats. I tot això en un marc en què no es discrimine cap alumne per tenir alguna deficiència o manca de capacitat. La diversitat té clarament un caràcter enriquidor.

Recentment s'ha elaborat un conjunt d'indicadors (Tony Booth i Mel Ainscow, 2002: *Guía para la evaluación y mejora de la educación inclusiva*) que, d'una banda, mesuren el grau d'inclusivitat d'una escola, i de l'altra, faciliten els canvis en un centre perquè s'hi adopte la inclusivitat com un principi de l'escola. L'organigrama que apareix tot seguit mostra com, a partir d'una variable general, es determinen les dimensions, i d'aquestes, els indicadors, i també la manera de mesurar-los. En aquest cas s'ha optat per preguntes binàries.



Quadre 6

1.5.4. Disseny de la investigació

Un colp s'han concretat els objectius, les hipòtesis, les variables i les unitats d'anàlisi, correspon decidir el disseny de la investigació: especificar com es realitzarà.

De la mateixa manera que les fases precedents, el disseny està determinat pels objectius d'estudi (d'exploració, descriptius, explicatius, predictius, avaluatius), els recursos (humans, materials, econòmics) i el termini de temps del qual es disposa per a materialitzar-lo. La valoració d'aquests factors determinarà la tipologia de disseny escollit. D'Ancona (1996), estableix tres tipologies de disseny: atenent el grau de compliment dels supòsits de l'experimentació (dissenys preexperimentals, quasiexperimentals i experimentals); atenent el tractament de la variable «temps» (dissenys transversals i longitudinals); atenent els objectius de la investigació (dissenys d'exploració, descriptius, explicatius, predictius i avaluatius).

Aquesta triple consideració incideix directament en la selecció d'una o de diverses estratègies d'investigació: ús de fonts documentals i estadístiques, estudi de casos, enquesta, experiment, anàlisi de dades i arxius, etc.

L'estratègia o estratègies finalment escollides influiran en:

- El disseny mostral: la mostra de l'estudi (individus, habitatges, entitats socials...), el seu volum i la forma de selecció.
- Les tècniques de recollida d'informació: revisió d'estadístiques i documents, observació experimental i observació directa simple, qüestionaris, entrevistes...
- Les tècniques d'anàlisi de dades: documentals, estadístiques, de contingut, estructurals...

Una vegada el disseny ha estat escollit, és necessari comprovar-ne la validesa. En cas que el disseny escollit s'adeqüe als tres elements esmentats amb anterioritat (objectius, temps i recursos), es passarà a analitzar uns altres criteris d'avaluació: la validesa interna, la validesa externa, la validesa de constructe i la validesa de conclusió estadística.

Tipus de disseny

Com s'ha esmentat amb anterioritat, existeixen diferents aspectes que permeten classificar els dissenys. Tot seguit es comentaran breument les principals característiques de cadascun.

Dissenys preexperimetals, quasiexperimentals i experimentals

Campbell i Stanley, en el llibre *Diseños experimentales* (1963), diferenciaren els tres tipus principals de dissenys d'investigació: els dissenys preexperimentals, els quasiexperimentals i els experimentals.

Abans de concretar i diferenciar els tres tipus de disseny, cal introduir un seguit de conceptes relatius a l'experimentació com a tècnica de recollida d'informació.

L'experimentació és una manera d'investigar basada en el control i la intervenció que realitza l'investigador sobre la realitat que estudia. Aquest control va encaminat a comprovar els efectes que té la variació de la variable que pot manipular l'investigador (variable independent) en l'ocurrència i les variacions de la variable que s'investiga (variable dependent).

Els components bàsics que defineixen l'experimentació són:

- *La manipulació experimental*. L'investigador pot manipular, abans de recollir la informació, les variables independents. És a dir, l'investigador crea una situació que li permet observar la influència causal d'una o diverses variables (independents) en una altra (dependent). La variable independent, de la qual es vol mesurar l'efecte, pot anar canviant de valors o de categories (estímuls experimentals) i es poden observar les variacions que es produeixen en la variable dependent.
- *El control de l'investigador*. La relació causal entre les variables que es tracta d'observar està condicionada al control dels factors que poden afectar-hi, com poden ser les variables estranyes. És per això que l'investigador haurà de controlar aquests factors.
- *L'aleatorització*. En qualsevol experiment hi ha, almenys, dos grups: el grup experimental que rep l'estímul experimental i l'efecte del qual es tracta de mesurar, i el grup de control que no el rep. Mitjançant la formació d'aquest segon grup s'intenta controlar l'efecte dels factors aliens a l'estímul experimental que poden influir en els resultats de l'estudi. És per això que ha de procurar-se la total equivalència inicial d'ambdós grups. Per a aconseguir-ho, és necessari que un dels criteris (o l'únic) per a assignar a cada individu a un grup, bé de control, bé experimental, siga l'atzar. El procés d'assignació doncs, ha de tenir un clar caràcter aleatori.

La distinció o classificació en aquests tres tipus de disseny (preexperimentals, experimentals i quasiexperimentals) respon al grau de compliment d'aquestes característiques fonamentals.

Els dissenys preexperimentals es caracteritzen per l'absència de manipulació de les variables que intervenen en la investigació, és a dir, l'investigador es limita a observar allò que s'estudia sense modificar res. Per mesurar una única vegada el fenomen, no es realitza una investigació paral·lela sobre un grup de control equivalent al grup experimental, excepte en les variables que es volen estudiar. Per a acabar, es caracteritzen per la falta de control de les possibles fonts d'invalidació de la investigació, de la seua validesa interna.

Malgrat que no es manipulen les variables, en aquest tipus d'investigacions sí que és possible trobar relacions causals a partir de l'anàlisi estadística multivariant de les dades, la qual permet esbrinar, un colp realitzada la recollida d'informació, les relacions de dependència i de correlació entre les variables.

Els dissenys preexperimentals inclouen tres modalitats: 1) disseny d'un grup experimental amb un sol mesurament (s'hi observa un grup en què s'ha fet incidir un estímul o variable experimental); 2) disseny d'un sol grup amb pretest i posttest (s'hi mesura el grup abans i després del tractament amb la variable experimental); 3) la comparació entre un grup experimental i un altre de control, sense cap mesura prèvia.

Com a exemple de disseny preexperimental, es poden destacar les enquestes usuals en què es realitza un únic mesurament de la realitat social després d'haver exposat el grup a una situació o d'haver-hi introduït un estímul. Les enquestes sobre el funcionament d'una normativa –com pot ser la llei del tabac– o les d'intenció de vot després d'unes declaracions intencionades des d'un partit polític, en són un bon exemple.

Cap dels tres dissenys compleix tots els requisits de l'experimentació i per aquesta raó molts autors els anomenen *dissenys preexperimentals*.

En els dissenys experimentals l'investigador efectua un seguit d'actuacions a priori, encaminades a controlar les possibles fonts d'invalidesa i fonamentades en el control i la intervenció de l'investigador en la realitat que analitza. És a dir, compleix totes les característiques esmentades amb anterioritat (la manipulació experimental, la formació de grups de control, l'assignació dels individus als grups (experimentals o de control)) han de ser totalment aleatòries.

Una vegada realitzat l'experiment, l'investigador compara els resultats en els diferents grups, per comprovar així l'efecte que ha tingut el valor atorgat a la variable independent. La comparació dels resultats grupals de la variable dependent sol realitzar-se mitjançant tècniques d'anàlisi de variància, les quals en permeten mesurar estadísticament les diferències de les mitjanes grupals. Si els resultats de l'anàlisi estadística conclouen que no existeixen diferències estadísticament significatives entre els valors de la variable dependent dels grups (els experimentals i els de control), no es podrà afirmar que el valor de la variable independent ha tingut efecte. Però si les diferències entre els grups són estadísticament significatives, s'esdevé que el valor que ha pres la variable independent ha produït un efecte en la dependent (en consonància amb les hipòtesis d'investigació).

Com a exemple, es pot considerar el següent: una empresa vol comprovar si un augment de la flexibilitat en l'horari de treball millora la productivitat. En concret, es tractaria d'analitzar si la flexibilitat de l'horari (variable independent) influeix en la productivitat del treballador (variable dependent). En un disseny experimental bàsic es confeccionarien dos grups de treballadors, un d'experimental i un de control. Els grups estarien formats per treballadors de característiques tan semblants com fóra possible, l'única diferència entre ells hauria de ser la manipulació de l'experiment: la flexibilitat de l'horari. Tots dos grups treballarien durant el temps establert en l'experiment i, posteriorment, s'analitzarien de la mateixa manera les productivitats dels treballadors d'ambdós grups.

Perquè les diferències observades foren atribuïbles a la variable independent, caldria comprovar prèviament la incidència d'unes altres variables, variables externes (característiques personals o socials dels participants en l'experiment, característiques contextuals...), que pogueren influir en la variable dependent, en aquest cas la productivitat.

Per a finalitzar, cal remarcar que els dissenys experimentals n'afavoreixen la validesa interna, és a dir, el control de les explicacions alternatives a les que s'analitzen en el fenomen que s'estudia. Tanmateix, se'n ressent la validesa externa; per tant, els resultats no es poden generalitzar a causa de la manipulació experimental i del baix nombre d'individus que generalment en formen la mostra (no superior a 200).

Els dissenys quasiexperimentals són una barreja de tots dos dissenys esmentats anteriorment. Així, s'hi poden realitzar-se –o no– manipulacions experimentals. Es diferencien dels experimentals en la distribució de la mostra d'estudi en els grups de control i experimental, que no es realitza aleatòriament. Pel que fa als preexperimentals, s'hi diferencien en l'estructuració de la situació que fa l'investigador de manera que en facilita l'anàlisi. No es limita únicament a l'observació.

Pel que fa a la classificació dels dissenys quasiexperimentals, la manipulació –o no– de la situació experimental, i l'existència –o no– dels grups de control i experimentals equiparables, són els trets que permeten agrupar-los.

Com a exemple, es pot considerar el mateix que en el cas anterior però amb modificacions. L'empresa podria seleccionar un conjunt de treballadors i mesurar-ne la productivitat en tres moments diferents. Posteriorment, se'ls flexibilitzaria l'horari de treball i se'ls tornaria a mesurar la productivitat en tres moments. Per a acabar, s'analitzarien totes les dades estadísticament. Aquesta repetició de les mesures abans i després del test, proporciona més poder d'aïllament dels efectes de la variable experimental.

Dissenys seccionals o transversals i dissenys longitudinals

Aquesta tipologia de dissenys pren el temps com a variable essencial.

Els dissenys seccionals o transversals es caracteritzen per recollir la informació en un únic moment en el temps. Aquests tipus de disseny no comprenen ni la diversitat d'observacions, ni de grups, ni tampoc de variables experimentals; queden limitats a una sola observació d'un sol grup en un sol moment. A causa de la simplicitat, aquesta classe d'experiments sol ser freqüent en les investigacions socials. S'hi empren tècniques de recollida de dades basades en l'observació directa, en l'enquesta, etc.

Com a exemple, es pot considerar l'anàlisi de la competència professional en un moment donat, dels treballadors d'una empresa, amb distinció del tipus de feina, del sexe dels treballadors, etc.

Per contra, els dissenys longitudinals es diferencien dels seccionals per recollir la informació en diversos punts temporals, amb la finalitat d'observar la dinàmica d'allò que s'està estudiant. La cronologia de les observacions està lligada a l'objecte d'estudi.

Per exemple, l'estudi de l'avaluació docent que realitza la Universitat Jaume I al professorat té un disseny longitudinal. La raó és que aquest s'avalua cada any, dues vegades al llarg del curs escolar i, a més, no hi ha cap variable experimental d'estudi. L'objectiu és bàsicament descriptiu.

Dissenys exploratius, descriptius, explicatius, predictius i avaluatius

La darrera tipologia de dissenys respon als objectius principals de la investigació. Cal dir que aquesta classificació dels dissenys no és mútuament excloent, és a dir, una mateixa investigació pot tenir objectius diferents segons la fase de la investigació en què es trobe. Cadascun dels dissenys ha d'anar en consonància amb els objectius corresponents. Així el disseny exploratori és l'adequat quan es pretén familiaritzar-se amb el problema d'investigació, té diverses finalitats: extraure variables importants i hipòtesis per a comprovar-les en indagacions posteriors, verificar la factibilitat de la investigació, comprovar quina estratègia o estratègies s'adeqüen més al projecte d'investigació, etc. El disseny descriptiu constitueix un pas previ a qualsevol investigació, abans d'intentar resoldre el problema cal descriure'l mitjançant alguna o diverses estratègies d'investigació (enquesta, estudi de casos, etc.). El disseny explicatiu té com a objectiu buscar possibles causes dels fets, accions, opinions o qualsevol fenomen que s'analitze. El disseny predictiu necessita alguns dels objectius anteriors. La finalitat és predir, o siga, trobar les variables que conjecturen la futura evolució de fenòmens concrets. Per a acabar, els dissenys avaluatius tenen per objecte comprovar l'adequació d'un programa o actuació respecte a les seues metes originals.

L'estudi de l'avalució docent del professorat de la Universitat té un objectiu primari bàsicament descriptiu: descriure la competència docent del professorat de l'UJI. Si s'intentaren analitzar les causes que produeixen el descontent en els treballadors d'una empresa s'utilitzaria un disseny explicatiu, en què es mesurarien les interrelacions i les influències de diverses variables.

La validesa del disseny

En definitiva, la validesa del disseny ha de ser el garant del compliment dels objectius marcats i de la representativitat dels resultats aconseguits respecte de la realitat social que estudien.

Per a aconseguir-ho, diferents autors, com ara Campbell, Stanley, Cook i Reichardt, proposaren quatre criteris d'avaluació dels dissenys d'investigació quantitatius:

La validesa interna es refereix a la concordança, dins de la mateixa investigació, dels resultats obtinguts i la realitat investigada. Així, el disseny ha de tenir en compte la possible intervenció de variables estranyes o alienes a les que constitueixen l'objecte d'investigació i que són, en moltes ocasions, la causa d'errors. Aquest tipus de variables poden considerar-se com a factors que influeixen en els resultats de la investigació i que, per tant, poden explicar-ne els resultats i deixar al marge les hipòtesis de partida. És a dir, si existeix la influència, no es podrà arribar en la investigació a una conclusió única. Variables ambientals de tipus físic o social, variables connexes amb les investigades o aquelles que fan referència a l'actitud de l'investigador, en són exemples.

Pel que fa al control d'aquestes variables, pot fer-se a priori (en el disseny de la investigació, sobretot en els dissenys experimentals en escollir els elements dels grups de control i experimental aleatòriament) o a posteriori (en el procés d'anàlisi de les dades mitjançant tècniques estadístiques multivariants).

Un exemple: en una investigació al voltant de la claredat d'un escrit propagandístic –mesurat per les respostes a un test–, volen considerar-se quines variables incideixen més en la consecució d'una bona qualificació en el test. De les diferents variables analitzades, s'observa la relació entre les variables «nivell d'estudis» (primaris, secundaris...) (variable independent) i «qualificació del test» (variable dependent). S'hi observa que la qualificació del test és més alta en aquelles persones amb més formació acadèmica.

Perquè aquesta relació siga vàlida, és necessari controlar l'efecte d'altres variables externes en la relació observada, com poden ser el coeficient intel·lectual, el coneixement del producte que es promociona en l'escrit, etc. Aquest control es podria fer a priori, seleccionant les persones a qui es realitzaria l'experiment, o a posteriori, utilitzant tècniques d'anàlisi multivariant. Com més variables externes es controlen, més alt serà el grau de validesa de la investigació.

Perquè es poguera concloure que existeix la relació entre les variables «nivell d'estudis» i «qualificació en el test», s'hauria de comprovar (prèviament o posterior) que, indistintament del coeficient intel·lectual o del grau de coneixement del producte, com més nivell d'estudis té la persona analitzada, més alta és la qualificació obtinguda en el test. En cas contrari, la relació no es podria afirmar.

La validesa externa al·ludeix a la concordança entre els resultats obtinguts en la investigació i la realitat d'una població diferent de la que ha estat estudiada, però que té característiques semblants. Per tant, el disseny ha de tenir en compte els factors que afecten la representativitat dels resultats i, en conseqüència, la possibilitat de fer-ne una generalització. La mostra, com a representació de la població, i el seu procediment d'elecció (preferiblement mitjançant tècniques de selecció aleatòries o probabilístiques), són els elements que cal controlar per a poder generalitzar els resultats.

Per exemple, els resultats d'una enquesta sobre la intenció de vot als subscrits a un periòdic de tendència conservadora no gaudeix d'aquesta validesa, ja que la mostra escollida no representa tota la població.

La validesa de constructe fa referència a l'elecció adequada dels indicadors com a representants de les variables abstractes, dels conceptes teòrics. És necessari parar molt de compte que els indicadors mesuren exactament els conceptes fonamentals de la investigació.

S'hi recomana utilitzar l'operacionalització múltiple (donar més d'una mesura per a cada concepte), ja que permet més aproximació al significat real del concepte. Posteriorment, i abans d'afirmar les conclusions, cal comprovar si realment s'ha mesurat el mateix concepte.

Per exemple, si la competència professional es mesura de tres maneres –coeficient intel·lectual, dedicació a l'empresa (hores de treball) i experiència en l'empresa–, els resultats no seran amb tota probabilitat els mateixos. La raó és que cada procediment mesura coses diferents i, per tant, les tres mesures emprades no mesuren el mateix concepte.

Per a acabar, la validesa de conclusió estadística es troba relacionada amb la tècnica d'anàlisi de dades emprada, amb la seua adequació i amb la fiabilitat dels resultats obtinguts. Així, caldrà estudiar si les dades compleixen una sèrie de requisits necessaris per a utilitzar un procediment estadístic determinat, si la mostra és suficientment gran per a poder generalitzar els resultats a la població, etc.

Per exemple, per a poder inferir a tota la població els resultats obtinguts d'una mostra representativa, és necessari comprovar que les dades obtingudes estan sotmeses a una determinada llei matemàtica.

1.5.5. La viabilitat de la investigació

El darrer component del projecte inclou l'exposició de les mínimes condicions perquè el projecte es puga dur a terme. Aquestes condicions es poden resumir en:

- Fonts: selecció d'obres clau i de bibliografia actualitzada.
- Recursos disponibles: recursos materials, humans i econòmics necessaris.
- Planificació del temps d'execució del projecte: cal establir l'ordre cronològic de les tasques. Això suposa delimitar la duració de cada fase d'investigació. Es poden utilitzar tècniques per al càlcul dels temps, com per exemple el mètode PERT.

En resum, a banda dels dos aspectes fonamentals, com són el temps i els recursos necessaris per a realitzar la investigació, és evident que els objectius determinen el disseny de la investigació. Segons la categoria a la qual pertanyen els objectius, el disseny serà un o un altre. L'elecció del disseny determina quina o

quines estratègies d'investigació cal escollir, les quals concreten les tècniques tant per a la recollida de la informació com per a la posterior anàlisi i interpretació. La investigació s'ha de dur a terme de manera que els resultats gaudisquen de la màxima validesa possible.

En últim lloc, i per a introduir el tema següent, se cita García Ferrando (1985):

> El paper de l'estadística en el procés d'investigació social està determinat amb suficient claredat. L'estadística s'utilitza per a operar amb números que reflecteixen valors de mesures, que se suposa que satisfan diferents supòsits. Les consideracions estadístiques s'introdueixen bàsicament en la fase analítica del procés d'investigació [...]. Si el problema d'investigació no està ben definit, de poc servirà la utilització d'un gran aparell estadístic, ja que els resultats no milloraran per això. L'estadística s'ha de considerar com un auxiliar imprescindible [...] és sempre una bona ajuda però mai un substitut per a un bon raonament teòric i un bon quefer metodològic.

TEMA 2

Introducció a l'estadística. Relació amb la investigació social

Objectius tema 2

- Conèixer els fonaments del procés estadístic.
- Saber ubicar el procés estadístic dins d'una investigació social.
- Comprendre i saber diferenciar els conceptes més bàsics de l'estadística.

2.1. Introducció

L'estadística, com es pot deduir del capítol anterior, està lligada a la investigació social des dels orígens. No obstant això, per si sola constitueix una branca científica. Si bé és cert que la investigació social hi ha estat lligada, no és menys cert que, com a part fonamental de les matemàtiques més aplicades, l'estadística ha anat evolucionant al llarg del temps i s'ha convertit en una eina indispensable per a moltes altres ciències.

El món del treball també és partícip dels avantatges que produeix una anàlisi quantitativa dels problemes als quals s'enfronta. Així, són moltes les empreses que prenen decisions basant-se fonamentalment en conclusions estadístiques; per exemple, un estudi de mercat pot motivar la innovació en un producte, una enquesta pot decidir si una empresa augmenta els horaris de compra en establiments comercials, o una enquesta de satisfacció laboral pot millorar la productivitat d'una empresa. De les anàlisis de les dades obtingudes en aquests tipus de recerques, es prenen les decisions oportunes i se solucionen els problemes. És important, doncs, que a les empreses hi hagi persones que coneguen els trets estadístics més elementals, ja que si bé les anàlisis estadístiques en profunditat són realitzades per especialistes, són els responsables de les empreses els qui han de saber quan és necessària l'aplicació d'aquests estudis, com es realitzen aproximadament i quines conclusions se'n poden obtenir.

En una investigació social, sobretot en les quantitatives, les tècniques estadístiques tenen una importància vital. D'una banda, permeten analitzar les dades obtingudes al llarg del procés d'investigació i, d'una altra, faciliten arguments per a determinar si les hipòtesis de partida són vàlides o no. A més a més, en algunes ocasions els resultats de les anàlisis de les dades provoquen que la investigació en canvie alguna de les hipòtesis inicials.

Per altra part, l'anàlisi estadística més comuna és aquella que normalment prové d'estudis que en moltes ocasions no pretenen crear teories en si, o que formen part d'una investigació molt més complexa. Aquests estudis solen ser de tipus quantitatiu. Altrament, aquests estudis consisteixen en l'anàlisi d'un conjunt de dades de variables quantitatives, i això és un procés que es realitza utilitzant diferents tècniques motivades pels objectius de l'anàlisi i per les dades. En essència, el primer que es fa quan s'han d'analitzar dades és representar-les de manera que en faciliten la comprensió. La part de l'estadística que se n'ocupa és l'estadística descriptiva. El pas següent és conèixer si els resultats descrits per les dades, que provenen d'una mostra, es poden generalitzar a tota la població amb un cert grau de confiança, és a dir, es pretén inferir els resultats. D'aquesta part s'ocupen l'estadística inferencial i la probabilitat.

En aquest capítol s'ubicarà l'estadística dins del procés d'una investigació social. A més a més, després de donar unes nocions sobre els orígens de l'estadística matemàtica, es mostrarà a grans trets el funcionament del procés estadístic més elemental.

2.2. Orígens

Fou Achenwall, professor de la Universitat de Gottingen, qui en 1748 introduí el terme *estadística* en el sentit actual. Va utilitzar-lo per a definir el tractament matemàtic que es dóna al procés de recerca de lleis sobre la regularitat de certs fenòmens socials.

Com s'ha comentat en la unitat anterior, al llarg de la història l'estadística ha sigut bàsicament descriptiva. Ha sigut utilitzada pels estats i per diferents corrents socials amb fins polítics, logístics i econòmics. Concretament, el seu origen es pot situar als inicis de les matemàtiques polítiques, en els segles XVII i XVIII, amb l'intent de trobar les lleis que regien els fenòmens socials. Cal assenyalar, però, que els camps d'activitat de les matemàtiques polítiques no tenen límits fixos, van des de la demografia fins l'economia, i no tan sols fent recomptes de dades i representant-les –objectiu de l'estadística descriptiva–, sinó traient conclusions de les dades i prenent decisions basades en els resultats obtinguts.

Per altra part, l'activitat estadística és essencialment matemàtica i, per tant, els seus principis han de provenir d'aquesta ciència, concretament del càlcul de probabilitats.

L'origen del càlcul de les probabilitats està en l'interès de les matemàtiques en els processos lligats a l'atzar. Se centra en el cavaller de Méré aquest primer impuls o crida a les matemàtiques. Aquest cavaller, un ric francès de mitjan segle XVII, que freqüentava les sales de joc en què s'apostaven nombroses quantitats de diners, va observar en diverses ocasions com la sort li era adversa. A més, no trobava l'explicació al fet que, jugant a obtenir almenys dues vegades dos sisos en llançar vint-i-quatre vegades consecutives dos daus, la sort li fóra desfavorable. Aquest fet i l'observació d'algunes regularitats en els resultats dels daus el dugueren a pensar que era possible l'existència d'una llei matemàtica que regia els resultats dels daus. Qui la descobrira podria realitzar les apostes amb més garanties d'èxit.

De Méré plantejà qüestions semblants a l'anterior a Blaise Pascal (1632-1662) el qual es comunicava epistolarment amb Pierre Simon de Fermat (1601-1665), i tots dos s'ocuparen de resoldre-les matemàticament. D'aquesta manera i sense adonar-se'n, estaven creant la disciplina matemàtica del càlcul de probabilitats. Fou l'holandès Christian Huygens (1629-1695) qui recollí el contingut de la correspondència entre Pascal i Fermat, així com les matemàtiques que hi subjauen, en l'obra *De ratiociniis in ludo aleae* (Sobre raonaments relatius als jocs de daus). Pot dir-se que l'obra del matemàtic holandès és la primera en què apareixen qüestions sobre probabilitat.

Importantíssims són els treballs del matemàtic suís Jacques Bernoulli (1654-1705) i dels seus continuadors, Abraham de Moivre (1667-1754) i Laplace (1749-1827), el qual va escriure l'obra *Théorie analytique des probabilités*, considerada la més

influent per a la posteritat en qüestions de probabilitat. Per la terminologia i pels resultats que inclou, és un llibre central en la bibliografia matemàtica del càlcul de probabilitats.

D'altra banda, cal recordar que a l'home de la segona meitat del segle XVIII no preocupaven tant els jocs d'atzar, sinó que les seues inquietuds relatives al tema que es tracta eren més àmplies: es preguntava la probabilitat que els habitants d'un determinat país sobrepassaren una determinada edat, que els beneficis d'una empresa superaren una determinada quantitat, que un grup social actuara d'una determinada manera, etc. Dos dels matemàtics del moment, Euler (1707-1783) i D'Alembert (1717-1783), escrigueren tractats sobre anualitats de capitalització, sortejos de loteria i esperança de vida.

L'estadística passà de ser una activitat aritmètica al servei de l'estat i un art de descriure en forma quantitativa, a ser una ciència que es preguntava per les causes i influències, tant naturals com polítiques i socials, d'esdeveniments, epidèmies, cataclismes o conductes socials.

L'estadística com a ciència, després de l'obra de Laplace, abandonà els mètodes merament descriptius i utilitzà com a fonament el càlcul de probabilitats.

Un dels representants més genuïns d'aquest nou mètode va ser el matemàtic, sociòleg i estadístic belga Adolphe Quetelet (1796-1874), que evidencià que és possible utilitzar l'estadística com a ciència capaç de calibrar els fenòmens socials.

Més endavant, l'antropòleg britànic Francis Galton (1822-1911) introduí la teoria de la correlació, imprescindible en tot procés estadístic.

En els albors del segle XX s'uniren definitivament, en tot procés estadístic, la descripció i anàlisi de dades, per una part, i el càlcul de probabilitats, per una altra. Tots dos han format el que actualment es coneix com a *estadística matemàtica*, la qual pretén situar en cada cas el fenomen observat o el procés en estudi, sota els pressupostos d'una llei matemàtica que permetrà extraure'n conseqüències i, posteriorment, les decisions més adequades, d'acord amb el tipus d'estudi que es realitze (matemàtic, econòmic, social, polític, etc.).

Lligat a aquesta nova concepció dels processos estadístics, aparegué Karl Pearson (1857-1936), amb qui el procés estadístic aplicat a fenòmens biològics i socials va experimentar un important auge. Una de les seues aportacions més importants és la distribució khi quadrat per a mesurar la bondat d'ajustament, que es va descobrir en 1900. En 1906 el rus Andrei Markov (1856-1922) inicià l'estudi dels esdeveniments encadenats, coneguts popularment com a *cadena de Markov*.

L'any 1931, el matemàtic Andrei N. Kolgomorov (1903-1987) establí una fonamentació axiomàtica del càlcul de probabilitats influït per la teoria de la mesura de Lebesgue. Llavors sorgí una ruptura entre l'anàlisi matemàtica clàssica (basada en l'estudi de les funcions contínues) i el càlcul de probabilitats, en el

qual és necessari estudiar processos discrets. La vinculació final d'aquestes dues disciplines matemàtiques es coneix millor a partir de la teoria de distribucions, que no és més que una generalització de la diferenciació, apareguda gràcies a Laurent Schwartz.

L'arribada de les computadores electròniques evidencià la relació entre el càlcul de probabilitats i els processos finits. També va provocar l'aparició de nous camps d'estudi, com ara la investigació operativa, la teoria de jocs, o la programació lineal.

Finalment, cal ressenyar el gran impuls actual dels processos estadístics, que interfereixen en les branques més modernes de les matemàtiques.

2.3. Tècniques estadístiques i procés estadístic

2.3.1. Procés estadístic

Com es dedueix de la introducció de l'epígraf anterior, històricament es diferencien dues nocions d'estadística diferents: una primera que s'utilitza fonamentalment per a descriure realitats mitjançant l'anàlisi de dades, i una segona que originàriament s'utilitzava per a realitzar apostes en els jocs d'atzar, i en fusionar-se amb l'estadística descriptiva, per a inferir resultats si les dades analitzades complien un conjunt de lleis.

Els tres termes emprats –*descriptiva*, *probabilitat* i *inferència*– són complementaris del procés estadístic més elemental i cadascun representa un tret característic dins del procés. No cal oblidar, però, que l'estadística realitza estudis sobre dades que per si soles proporcionen una informació molt incompleta. Per exemple: una gran multinacional ha comunicat a tots els gerents de les empreses que la formen, que realitzen una anàlisi del sou dels treballadors, així com de les hores de treball. Per aquesta raó, els gerents comuniquen al personal administratiu que els presenten les llistes on apareguen les dades dels treballadors, les hores de treball i el sou mensual que reben. Amb aquest recull de dades, sobre 125 treballadors, el gerent difícilment pot extraure conclusions, les dades s'han de manipular per a fer productiva l'anàlisi.

	Hores de treball			
Sou (milers €)	[25, 28)	[28, 31)	[31, 34)	.....
[10, 12)	10%	15%	.....	
[12, 14)	20%	.....	.....	
[14, 18)	.....	.....	.....	
.....	.....	.....	.....	

Anàlisi de les variables «sou» i «hores de treball» conjuntament

Quadre 1

Sou (milers €)	Percentatge
[10, 12)	12%
[12, 14)	25%
.....	.....

Anàlisi de la variable «sou»

Quadre 2

Així, en primer lloc caldria que l'encarregat de realitzar aquest estudi definira amb total exactitud les magnituds que s'estudiaran: si es consideren o no les hores extra, les primes per productivitat, si el sou que s'estudia és el brut, etc. Un colp especificat allò que es pretén analitzar, demanaria les dades adequades i en començaria l'anàlisi. D'aquesta manera, es podria construir una taula (quadre 2) en què aparegueren els percentatges de treballadors que cobren sous similars i treballen aproximadament les mateixes hores. Si en la taula, a més a més, hi haguera un ordre, llavors la informació seria molt més clara. També es podrien construir taules semblants per a cada una de les dues variables de manera individual (quadre 1). L'estadística facilita unes normes per a construir-les.

Si es desitja aprofundir en l'anàlisi, l'estadística permet l'aplicació de diferents procediments estadístics, que en moltes ocasions es poden realitzar amb programari informàtic: confeccionar gràfics per a cada variable o conjuntament, calcular salaris mitjans per categories, esbrinar el grau d'equitat de la distribució dels salaris, conèixer si hi ha cap relació important entre hores de treball i salari, i moltes altres possibilitats.

Com es pot observar, fins ara els únics procediments esmentats proporcionen una descripció del recull de dades, és a dir, mostren la informació que dilueixen les dades. D'aquesta tasca de descripció de les dades, se n'ocupa l'estadística descriptiva.

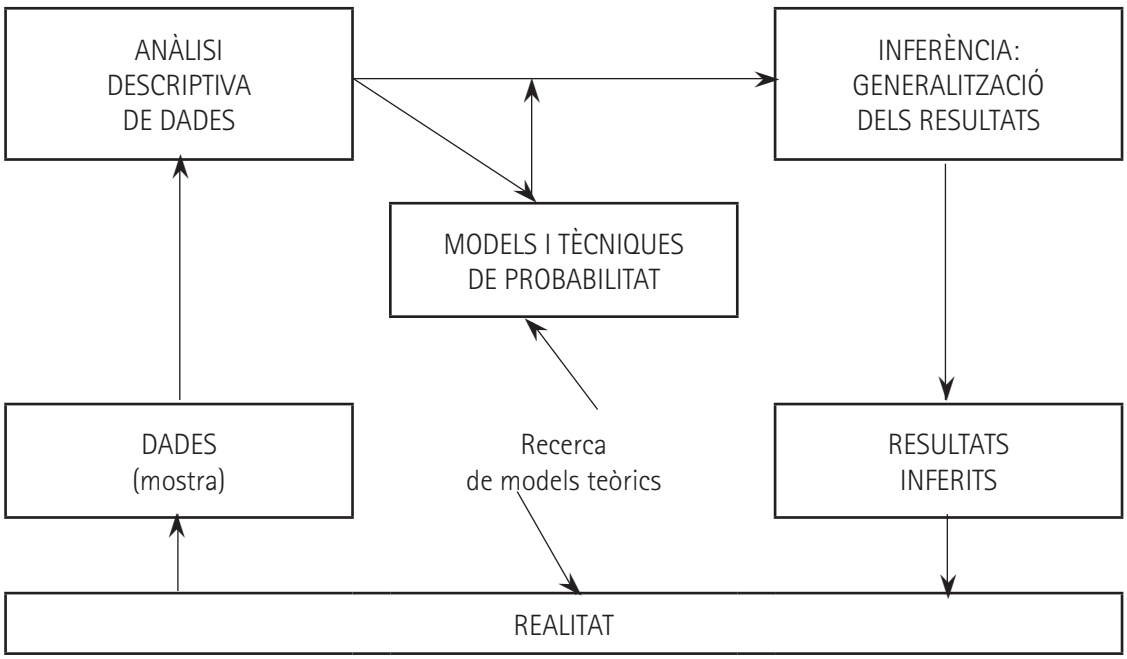
Per altra part i seguint amb l'exemple, és possible que la multinacional decidisca realitzar l'estudi per a tots els treballadors de la multinacional (120.000) i designe un únic responsable per a realitzar el treball. En aquest cas, per raons tant econòmiques com d'eficiència, no és necessari analitzar totes les dades. Fent ús d'algunes tècniques estadístiques es poden obtenir uns resultats molt pròxims a la realitat analitzant un nombre molt menor de dades. La inferència estadística és l'encarregada de proveir els instruments necessaris per a inferir resultats vàlids per a tota la població a partir dels obtinguts per a un conjunt de dades (la mostra). Per tant, tots aquells continguts que facen referència a l'elecció de quines i quantes dades de la població formaran la mostra, del control de l'error de les estimacions, de les previsions per a anys futurs, del grau de dependència d'unes variables en comparació amb unes altres per a tota la població, etc., pertanyen a aquesta branca de l'estadística.

Hi ha una altra qüestió que cal tenir present: els fonaments de la inferència estadística són resultats teòrics, extrets de models teòrics de caire matemàtic que formen part de la branca de l'estadística anomenada *probabilitat*. Els continguts d'aquesta part s'han obtingut a partir de regularitats observades en la realitat i el posterior procés de modelatge matemàtic. D'aquesta manera, donat un conjunt de dades, és possible trobar un model teòric, anomenat *model de probabilitat*, que s'adeqüe a la distribució de les dites dades. Aquest fet permet teoritzar-les i, en conseqüència, inferir les característiques del model teòric a la resta de dades de la població. És evident, doncs, la importància d'escollir una mostra representativa.

Reprenent l'exemple, l'encarregat de l'estudi haurà d'escollir la mostra i fer-ne l'estudi descriptiu. Després, caldrà que esbrine el model teòric que més s'aproxime al model descrit per les dades. Una vegada realitzada aquesta tasca, podrà emprar els procediments probabilístics que li permetran inferir els resultats amb un determinat grau de certesa. Els tests d'hipòtesi i la creació d'intervals de confiança en són una mostra.

Pel que fa a l'anàlisi conjunta de més d'una variable, el procés estadístic descrit també permet estimar, amb un cert control de l'error d'estimació, tant la relació existent entre les variables com les prediccions dels valors d'unes variables en funció dels valors que prenen unes altres.

L'exemple tan senzill que s'està tractant al llarg d'aquest punt reflecteix el procés estadístic més usual, així com les interrelacions existents entre els seus tres components més importants: l'estadística descriptiva, la probabilitat i la inferència estadística. El gràfic següent (quadre 3) resumeix el procés estadístic.



Quadre 3

2.3.2. Tècniques estadístiques d'anàlisi de dades

Per *tècniques estadístiques* es pot entendre un conjunt molt ampli de procediments estadístics, però en tots, els elements primordials són les variables, les dades i els objectius que han propiciat l'estudi estadístic. De fet, són aquests darrers els que determinen vertaderament quines tècniques són les escollides per a aconseguir les finalitats preteses. D'aquesta manera, no són necessaris els mateixos procediments si es desitja predir el volum de vendes d'una empresa en un mes determinat, basant-se en les vendes d'anys anteriors, que si es volen analitzar les cinquanta preguntes d'una enquesta sobre l'envàs d'un producte, o predir els resultats d'unes eleccions a partir de l'anàlisi d'una enquesta d'opinió.

Així, si únicament s'estudia una variable quantitativa es durà a terme anàlogament el procdiment esmentat anteriorment; se seguirà el procés estadístic elemental: de la descripció a la inferència. Tècniques basades en els tests d'hipòtesi, construccions d'intervals de confiança o tests no paramètrics, són exemples de procediments emprats en l'anàlisi d'una variable. Tanmateix, si en són dues les estudiades, s'hi podrà conèixer també, amb un cert grau de confiança, el tipus de relació existent entre les dues variables (lineal, quadràtica, exponencial...). D'aquesta manera es podran predir els valors d'una variable a partir dels valors d'unes altres. Especial importància té si una de les variables estudiades és el temps. En aquest cas s'empraran tècniques basades en les sèries temporals, que permetran realitzar prediccions futures.

Si s'estudien més de dues variables, a més a més de conèixer-ne les relacions existents entre si, les possibles prediccions d'unes en funció d'unes altres, i altres procediments estadístics més tècnics, és possible la classificació de les dades en conjunts atenent criteris de semblança o de significació. L' anàlisi clúster n'és un bon exemple.

Si s'estudien dades de tipus qualitatiu, també es poden usar tècniques estadístiques, amb les quals és possible, per exemple, representar gràficament o mitjançant taules estadístiques, dades referents a una o més variables qualitatives, conèixer el grau d'interdependència d'un conjunt de variables qualitatives, classificar les dades en grups basant-se en criteris de semblança, etc.

Cal dir, a més, que existeixen moltes eines informàtiques per a dur a terme quasi tots els procediments estadístics. Per aquesta raó, en l'anàlisi de les dades sol ser més important allò que es vol saber que com fer-ho.

Per a finalitzar aquest punt, cal remarcar que a causa del caràcter introductori del text des del punt de vista de l'estadística, únicament es realitzarà l'estudi del procés estadístic esmentat anteriorment i de les tècniques estadístiques més freqüents que té associades.

2.4. L'estadística en una investigació social

En una investigació social, fonamentalment en aquelles de caire quantitatiu, les tècniques estadístiques s'empren principalment en el procés d'anàlisi de la informació recollida i en el disseny de la mostra. Així, després de realitzar el treball de recerca sobre allò que s'està investigant, de tenir-ne clars els objectius, d'haver operacionalitzat les variables generals i construït les hipòtesis, s'escullen el disseny i l'estratègia d'investigació que s'emprarà, la qual determinarà, entre altres coses, la mostra i les tècniques de recollida de la informació.

L'estratègia no estableix únicament com s'ha de recollir la informació, sinó també quines tècniques estadístiques cal emprar per a analitzar les dades. Així, si únicament s'està estudiant una variable quantitativa, com pot ser el sou dels valencians en finalitzar l'any 2005, cal emprar tècniques d'anàlisi univariant. En aquest cas el disseny que cal utilitzar és el seccional; la tècnica de recollida d'informació, l'enquesta, i l'anàlisi estadística ha de seguir el procés estadístic elemental; en primer lloc, descripció de les dades que formen la mostra i, en segon lloc, la inferència dels resultats per a tota la població. Amb posterioritat, s'han de comparar els resultats obtinguts amb les hipòtesis. Tanmateix, si el disseny és experimental i s'hi estudien diferents variables, caldrà emprar tècniques d'anàlisi multivariant de dades per a poder traure'n resultats. Per exemple, si sobre un grup d'adults es vol conèixer l'efecte que té una campanya publicitària segons el temps durant el qual hi està exposat un individu (1 hora o 1,5 hores), i segons el tipus de transmissió d'informació (visual, escrita...), cal emprar aquestes tècniques per a conèixer les relacions causals de les variables independents –en aquest cas, el temps i el mode de transmissió–, i la variable dependent –l'efecte que produeixen.

Per altra part, les tècniques estadístiques són decisives a l'hora d'acceptar o refutar les hipòtesis de la investigació. Com que la investigació és de caire quantitatiu, les hipòtesis, que relacionen variables, han de ser constatades empíricament. És per això que l'anàlisi de les dades es pot considerar resultat empíric de les variables i permet argumentar la validesa de les hipòtesis o, per contra, el seu error.

D'altra banda, aquesta no és l'única aportació a la investigació de les tècniques estadístiques. De fet, els resultats obtinguts en els processos estadístics poden modificar algunes qüestions inicials de la investigació. És a dir, poden originar noves hipòtesi que en un principi no es contemplaven o no havien sorgit del procés heurístic de l'investigador. Així doncs, la investigació tornaria a començar, aquesta vegada amb noves hipòtesis que caldria verificar.

El gràfic que apareix tot seguit (quadre 4) sintetitza el que hem exposat en aquest apartat.

REALITAT
REALITAT
REALITAT
REALITAT
PROBLEMA QUE CAL INVESTIGAR
HIPÒTESIS DE RESOLUCIÓ
DISSENY DE LA INVESTIGACIÓ
OPERACIONALITZACIÓ
VERIFICACIÓ
DADES
DESCRIPCIÓ
ANÀLISI ESTADÍSTIC: PROCÉS ESTADÍSTIC I TÈCNIQUES ESTADÍSTIQUES
PROBABILITAT
CONCLUSIONS DE LA INFERÈNCIA ESTADÍSTICA
CONTRAST D'HIPÒTESIS
GENERALITZACIONS EMPÍRIQUES

Quadre 4

TEMA 3

Distribució estadística d'una variable (I): taules i gràfics

Objectius tema 3

- Conèixer els conceptes bàsics de les variables estadístiques.
- Saber classificar les variables estadístiques.
- Saber analitzar i realitzar taules de freqüències d'un conjunt de dades.
- Conèixer les diferències entre les taules de dades sense agrupar i les taules de dades agrupades.
- Conèixer els conceptes més bàsics de l'estadística.
- Saber interpretar i construir els principals gràfics estadístics.

3.1. Introducció

A l'hora de realitzar un estudi estadístic, el coneixement de les dades sense cap tipus d'anàlisi és per si mateix del tot insuficient en el moment d'obtenir-ne informació. En realitat, si el nombre de dades és relativament gran, la informació que aquestes amaguen es perd entre la multitud de valors. Per exemple, si considerem el conjunt de valors referents a les puntuacions obtingudes per 200 aspirants en una prova per a obtenir un lloc de treball en un banc (quadre 1), molt poques conclusions en podem traure, no podem saber quants participants han aprovat, ni quines han estat les notes més repetides, ni tan sols és possible fer-se una idea de com ha anat la prova. Per a poder descriure les dades i així fer-ne una anàlisi, és necessari realitzar certes operacions sobre les dades, de manera que siguen més comprensibles. Les taules i els gràfics estadístics són els elements descriptors més elementals i que amb assiduïtat apareixen en els mitjans de comunicacions. Si sobre el conjunt de valors del quadre 1 fem un senzill recompte i construïm una taula de freqüències que el mostre, s'obtenen les dades del quadre 2.

4	5	7	6	5	4	6	5	4	6	6	6	7	5	6	4	4	4	4	8
6	4	5	4	7	5	8	7	6	4	5	5	4	8	5	6	6	5	4	5
6	6	5	5	7	5	7	6	5	4	5	6	5	6	5	6	5	5	5	6
7	6	5	6	6	5	5	5	5	6	5	5	7	6	5	5	7	5	6	4
5	5	6	5	5	8	4	6	4	5	6	5	4	6	5	6	7	7	6	5
5	8	6	6	6	5	5	6	8	6	5	5	8	6	6	6	7	6	5	6
4	4	7	6	4	5	6	5	8	5	6	6	6	5	3	7	8	6	8	5
6	8	6	4	5	4	7	5	7	5	3	4	7	5	8	7	5	4	6	5
7	6	5	6	4	6	6	7	5	4	7	7	6	7	7	2	4	5	6	3
5	3	6	4	6	6	6	4	8	4	5	7	5	7	5	7	7	6	7	5

Quadre 1

Puntuacions	Freqüència amb què apareixen	Freqüència amb què apareixen en %
2	1	0,5
3	4	2
4	30	15
5	64	32
6	59	29,5
7	29	14,5
8	13	6,5

Quadre 2

Com es pot observar, aquesta taula és molt més clarificadora, ja que és possible fer-se una idea de com estan distribuïdes les dades. Així, s'hi observa que la prova ha estat superada per la majoria dels aspirants (un 78,5%) encara que les notes no han sigut massa bones: tan sols un 6,5% ha obtingut una qualificació superior a 7.

D'altra banda, també és possible realitzar una sèrie de gràfics que resumisquen les dades d'una manera més gràfica. A la figura 1 n'apareixen dos tipus per al conjunt de valors del quadre 1.

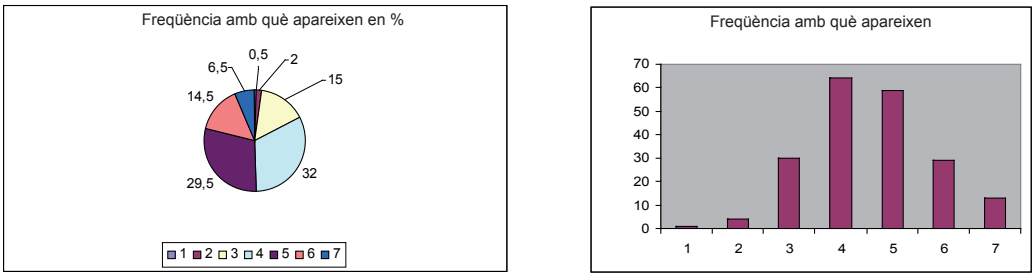


Figura 1

Al llarg d'aquesta unitat es realitzarà una introducció a aquestes dues eines estadístiques.

3.2. Conceptes preliminars

En les unitats anteriors ja s'han comentat dos dels conceptes fonamentals de l'estadística, com són la població d'estudi i la mostra. No obstant això, és convenient recordar les definicions i afegir-ne alguna de relacionada:

Població: És el conjunt d'elements o d'individus subjectes a estudi i dels quals es vol obtenir un resultat.

Exemple: Si es volen conèixer els ingressos dels 40 milions de persones de l'estat espanyol, aleshores els 40 milions d'ingressos constitueixen la població objecte d'estudi.

Paràmetre: És una mesura descriptiva de la població total, de totes les observacions.

Exemple: La mitjana aritmètica dels ingressos o l'ingrés que apareix reflectit més vegades en les dades de la població.

Mostra: Conjunt d'elements que formen part de la població total.

Grandària de la mostra: És el nombre d'elements o observacions que formen la mostra.

Estadístic: És una mesura descriptiva de la mostra, que estima el paràmetre de la població.

Exemple: La mitjana aritmètica de les dades que formen la mostra és un exemple d'estadístic.

3.2.1. Variables estadístiques

Quan s'estudia el comportament estadístic d'una característica que presenten els individus d'una població o d'una mostra, el primer que és necessari realitzar és mesurar-la. És a dir, cal associar a cada individu un nombre que permeta quantificar les diferents modalitats del caràcter. Si açò no és possible, cal classificar o ordenar els elements de la mostra segons les diferents modalitats que pot prendre el caràcter que s'observa. Així, si el que s'està estudiant és el nombre de vehicles que tenen les famílies espanyoles, és simple associar un nombre enter a cada element de la mostra: tindran 0, 1, 2, 3 o més vehicles. Tanmateix, si sobre la mateixa mostra s'estudia la població de residència habitual de cada família, les dades únicament es podran classificar per poblacions (Borriana, Castelló, Vila-real...). A més a més, si allò que es vol estudiar és el grau de satisfacció (baix, regular, bé, molt bé) que té cada família amb els serveis que li presta la Comunitat Valenciana, les modalitats seran susceptibles d'ordre.

Les modalitats que pren un caràcter, cal que estiguen ben definides, i que siguen exhaustives i mútuament excloents, és a dir, s'ha de saber amb total exactitud quines són les modalitats que presenta el caràcter i com reconèixer els individus que en manifesten una o una altra. A més, tots els individus en relació a aquest caràcter han de presentar-ne almenys una, i només una. Per exemple, si sobre el professorat de la universitat s'està estudiant el nombre d'articles publicats, cada professor en tindrà un nombre determinat i només un. Si, per contra, s'estudia el seu nivell d'empatia i les modalitats són «poca», «regular» o «molta», a cada professor únicament es podrà associar una de les tres modalitats.

Variables qualitatives i quantitatives

Les variables es poden classificar també, tal com s'està veient, segons la seua naturalesa. Així, no es pot analitzar de la mateixa manera la formació acadèmica de l'alumnat de l'UJI, que el nombre de crèdits que aquest té aprovats. Totes dues variables tenen naturalesa diferent. Mentre que en la primera les modalitats no es poden mesurar, en la segona sí que és possible fer-ho. Tanmateix, en algunes variables que no prenen valors numèrics, les modalitats sí que admeten una gradació. Per exemple, la variable «posició» que un individu pren enfront de l'Estatut d'Autonomia de Catalunya pot tenir les categories «molt d'acord», «d'acord», «en contra», «molt en contra», les quals són susceptibles de gradació.

Les variables en què únicament és possible un recompte del nombre d'elements de la població o mostra que posseeix una de les seues modalitats s'anomenen *variables qualitatives* o *atributs*. Les modalitats d'aquest tipus de variables ni tan sols admeten una gradació, i menys encara una mesura numèrica. Són variables com ara el sexe d'una persona, la confessionalitat, etc. Les modalitats que poden prendre s'anomenen *categories*. Així, les categories de la variable «sexe» són «masculí» i «femení».

La resta de variables en què, a més del recompte del nombre d'elements de la població o mostra que posseïx una de les seues modalitats, també és possible assignar una mesura a la mateixa modalitat, s'anomenen *variables quantitatives*. Són, per exemple, el pes, l'alçada, el sou mensual, el grau de duresa, etc.

Aquestes darreres variables, les quantitatives, també poden classificar-se en *discretes* i *contínues*. Una variable contínua és aquella que pot prendre qualsevol valor dins d'un rang donat. Independentment de la proximitat de dues observacions, si l'instrument de mesura és suficientment precís, sempre es podrà trobar una tercera observació entre les dues primeres. Per exemple, entre dues persones de 178 cm i 178,35 cm d'alçada, respectivament, pot trobar-se una persona de 178,25 cm d'alçada.

Una variable discreta està limitada per a certs valors, generalment nombres enters. Es diferencien de les quantitatives que, donades dues observacions suficientment pròximes, no es pot trobar cap observació entre totes dues. En són exemples, el nombre de fills de les famílies, el nombre de vehicles que tenen les empreses, el nombre de turistes que visiten un país, etc.

La variable «estadística» es denota per les majúscules. Així mateix, cada una d'aquestes variables pot prendre diferents valors, la notació dels quals és la següent:

$$X = (x_1, x_2,\ x_3, \ldots\ldots, x_{k-2},\ x_{k-1},\ x_k)$$

Escales de mesura

El concepte natural de mesurar una qualitat consisteix a assignar un nombre a les diferents modalitats que presenta. Així, les variables quantitatives es poden mesurar directament i, en canvi, en les qualitatives no té cap sentit empíric atribuir un valor numèric, és a dir, no són mesurables. Malgrat tot, sí que és factible ampliar el concepte *mesura* per a, d'aquesta manera, poder mesurar qualsevol variable.

Mesurar els resultats d'una variable consisteix a assignar-los un nombre de forma unívoca, fet que implica fixar escales i regles de mesura que dependran de la naturalesa de les observacions. Usualment es reconeixen quatre nivells de mesura, que de menys a més gradació són: nominal o cardinal, ordinal, d'interval i de raó o proporció.

Escala nominal: el nivell més baix de mesura és aquell en què únicament s'utilitzen els nombres per a identificar cada modalitat de la variable. D'aquesta manera, tots els elements de la població o la mostra que presenten la mateixa modalitat tindran assignat el mateix nombre. Aquesta és la regla de mesura que presenta. Per exemple, quan s'analitza una mostra de persones respecte a l'afinitat política, s'atribueix un 1 a les persones que simpatitzen amb el PSOE, un 2 a les del PP, un 3 a les d'EU, un 4 a les de CiU i un 5 a les d'altres afinitats. La finalitat d'aquesta escala és, doncs, classificar o enumerar.

Escala ordinal: el nivell de mesura ordinal és aquell en què els elements poden, a més a més de comparar-se per a comprovar si són o no iguals, ser ordenats de menys a més (d'agradar molt a no agradar gens, de difícil a fàcil, etc.) i en què no es pot establir cap relació més. D'aquesta manera, es pot saber que un element és més gran que un altre però no quant. Els nombres en aquesta escala indiquen l'ordre dins dels elements als quals s'aplica, i la regla de mesura d'aquesta escala és: a un element que siga més petit que un altre, li correspondrà un nombre també inferior. Per exemple, la variable «grau de satisfacció sobre la feina», pot prendre el valor 1 si és molt dolent, 2 si és dolent, 3 si és regular, 4 si és bo i 5 si és molt bo.

Escala d'interval: és el nivell de mesura immediatament superior a l'ordinal. En aquesta escala no sols es poden classificar i ordenar les dades, sinó que també és factible calcular la distància que les separa. Així com en les dues primeres escales únicament calia conèixer el valor associat a la dada, en aquest cas són necessaris un origen i una unitat de mesura per a determinar la distància entre dues dades. Per exemple, el caràcter temps, segons el calendari, segueix aquesta escala.

L'escala d'aquest nivell parteix d'un origen arbitrari i d'una unitat també arbitrària de mesura, encara que, una vegada fixada, ambdós romanen invariables en el transcurs dels mesuraments. La regla d'aquesta escala és: a observacions amb la mateixa quantitat d'unitats empíriques de diferència a l'origen, els ha de correspondre el mateix nombre. En aquesta escala, *mesurar* significa 'atribuir nombres que proporcionen distàncies entre les observacions'. Per exemple, l'escala centígrada té com a origen arbitrari la temperatura de fusió del gel, i com a unitat de mesura, també arbitrària, el grau centígrad. En aquest darrer exemple no existeix res concret que haja pogut obligar a fixar la temperatura de zero graus, sinó que simplement és un punt de referència arbitrari. S'hauria pogut establir l'escala de manera que el zero fóra un temperatura més calenta.

Escala de raó: en aquest nivell de mesura es poden establir les mateixes relacions empíriques pròpies del nivell de l'interval. També té un origen i una unitat de mesura arbitrària. La diferència amb l'escala d'interval és que en l'escala de raó l'origen és significatiu, és a dir, té un origen natural o absolut. Per exemple, caràcters com la longitud, el volum, la rendibilitat, etc., tenen un zero natural i, per tant, un nivell de mesura de raó.

Cada nombre assenyala les unitats de mesura que hi ha entre l'origen absolut i el nombre. La regla de mesura és: les observacions que estiguen a la mateixa distància de l'origen absolut tenen assignat el mateix nombre. L'exemple clàssic d'aquesta escala és el de les temperatures absolutes, que tenen per unitat de mesura el grau Kelvin.

3.3. Taules de freqüències

Abans de construir les taules de freqüències, cal realitzar una sèrie de definicions. Per a aclarir els conceptes s'exemplificaran les definicions segons el següent recull de 20 dades corresponents al nombre de telefonades enregistrades en diferents empreses entre les 10 h i les 10.30 h:

15, 5, 10, 5, 5, 6, 5, 6, 5, 6, 7, 10, 10, 12, 11, 11, 12, 15, 12, 15

Cal recordar, però, que els diferents valors que pot prendre la variable estadística es denoten mitjançant x_i. En aquest cas, ordenant-los de més petit a més gran, $X_1 = 5,\ X_2 = 6,\ X_3 = 7,\ X_4 = 10,\ X_5 = 12,\ X_6 = 15$.

S'anomena *freqüència absoluta del valor* x_i el nombre de vegades que apareix repetida l'observació en el recull de dades. Es representa per n_i. En l'exemple, la freqüència absoluta del valor x_2 és 3 ($n_2 = 3$), ja que la dada 6 es repeteix dues vegades en el conjunt de les dades de la mostra.

S'anomena *freqüència relativa del valor* x_i el quocient entre la freqüència absoluta de x_i i el nombre total de dades *n*. Es representa per f_i i, evidentment, és la proporció en què es troba el valor x_i dins del conjunt de dades en tant per u; $f_i = \frac{n_i}{n}$. En l'exemple, $f_2 = \frac{n_2}{n} = \frac{3}{20} = 0{,}15$. Per tant, el 15% de les dades són sisos.

D'altra banda, suposant que es disposara de *k* dades diferents, es compleix que la suma de tots els n_i és *n* ($n_1 + n_2 + ... + n_k = n$), i també que la suma de les freqüències relatives és igual a la unitat ($f_1 + f_2 + ... + f_k = 1$).

S'anomena *freqüència absoluta acumulada del valor* x_i el nombre de dades del recull que són inferiors o iguals a x_i. Es representa per N_i i el seu valor es calcula a partir de les freqüències absolutes; $N_i = n_1 + n_2 + ... + n_i$ (assumint que $x_1 < x_2 < ..... < x_i$). En l'exemple, $N_2 = 8$, ja que hi ha cinc cincs i tres sisos.

S'anomena *freqüència relativa acumulada del valor* x_i el quocient entre la freqüència absoluta acumulada de x_i i el nombre total de dades *n*. Es representa per F_i i, evidentment, és la proporció en què es troben els valors inferiors o iguals a x_i dins del conjunt de dades en tant per u; $F_i = \frac{N_i}{n}$. En l'exemple, $F_2 = \frac{N_2}{n} = \frac{8}{20} = 0{,}4$. Per tant, el 40% de les dades són sisos o valors inferiors a sis. També hi ha una altra manera de calcular F_i a partir de les freqüències relatives, ja que $F_i = f_1 + f_2 + ... + f_i$ (assumint que $x_1 < x_2 < ..... < x_i$).

Les freqüències acumulades també compleixen dues propietats trivials com a conseqüència de les seues definicions: suposant que es disposara de *k* dades diferents, es compliria que $N_k = n$ i $F_k = 1$.

És important remarcar que per a calcular freqüències acumulades és necessari que les variables que s'estudien siguen ordenables, és a dir, ha de ser possible establir una relació d'ordre entre els valors de les variables. Tanmateix, en les variables no ordenables no té sentit realitzar els càlculs esmentats.

Aquestes definicions permeten resumir les dades. No obstant això, la manera més adequada per a sintetitzar les dades és mitjançant el que es denomina *taula de freqüències*, on apareixen distribuïdes les dades segons les freqüències. Al mateix temps reflecteix tots els conceptes esmentats amb anterioritat. En l'exemple:

x_i	n_i	N_i	f_i	F_i
5	5	5	0,25	0,25
6	3	8	0,15	0,40
7	1	9	0,05	0,45
10	3	12	0,15	0,60
11	2	14	0,10	0,70
12	3	17	0,15	0,85
15	3	20	0,15	1
	20		1	

Taula 1

Cal remarcar que les dades estan ordenades (taula 1) (en aquest cas en ordre creixent) per a possibles actuacions posteriors sobre la taula, com pot ser el càlcul de la mediana, dels quartils, etc. (com es comprovarà amb posterioritat).

Dades agrupades

En ocasions el nombre de dades diferents que s'està estudiant és molt nombrós. Llavors, si es decidira construir una taula com l'anterior, la columna relativa a les x_i seria molt extensa, només cal pensar, per exemple, en dues-centes dades diferents dins d'un recull de quatre-centes.

La solució a aquesta qüestió consisteix a agrupar les dades en intervals o classes, de manera que cada dada pertanga a un –i sols un– interval. En conseqüència, els conceptes relatius a la freqüència que fins ara es referien als valors diferents de les dades, en realitzar l'agrupació han de fer referència als intervals.

Aquesta pràctica, malgrat que ajuda a resumir i clarificar la informació, té un inconvenient: es perd informació sobre la mateixa distribució de dades. En agrupar-les en els intervals, els valors reals es difuminen.

Un interval se sol representar[1] per $[L_{i-1},\ L_i)$ i es defineix com el conjunt format per tots els valors reals que són superiors o iguals a L_{i-1} (extrem inferior) i inferiors a L_i (extrem superior). Així, per exemple, l'interval [3,25 , 4,15) està format per tots els nombres reals superiors o iguals a 3,25 i, al mateix temps, inferiors a 4,15. Cal remarcar que el valor 4,15 no pertany a l'interval.

Pel que fa al concepte *freqüència absoluta d'un interval*, si s'haguera decidit agrupar les dades de l'exemple que s'ha estat considerant al llarg d'aquest epígraf i [10, 12,5) fóra un dels intervals, s'establirien les freqüències reflectides en la taula 2:

$[L_{i-1}, L_i)$	n_i	N_i	f_i	F_i
[10, 12,5)	8 (3 deus, 2 onzes i 3 dotzes)	17 (Hi ha 17 valors més petits que 12,5)	$\frac{8}{20} = 0{,}4$	$\frac{17}{20} = 0{,}85$

Taula 2

Es donen ara algunes definicions relatives als intervals, que s'utilitzaran al llarg del text. En totes les definicions es considerarà l'interval [10, 12,5) per a exemplificar els conceptes i així facilitar-ne la comprensió.

S'anomena *marca de classe* la mitjana aritmètica dels dos extrems de l'interval. És, evidentment, el valor central de l'interval, ja que equidista dels extrems. Es denota per c_i. Es calcula $c_i = \frac{L_{i-1} + L_i}{2}$. En l'exemple, $c_i = \frac{10 + 12{,}5}{2} = 11{,}25$.

S'anomena *amplària d'un interval* la distància que hi ha entre els extrems. Es denota per a_i i es calcula $a_i = L_i - L_{i-1}$. En l'exemple, $a_i = 12{,}5 - 10 = 2{,}5$.

S'anomena *densitat de freqüència absoluta d'un interval* el quocient entre la freqüència absoluta de l'interval i la seua amplària. Es denota per d_i. Es calcula $d_i = \frac{n_i}{a_i}$ i en l'exemple, $d_i = \frac{8}{2{,}5} = 3{,}2$.

Un colp establertes aquestes definicions és factible respondre a la pregunta que un es planteja si decideix agrupar les dades: quants intervals s'han de prendre i com s'han de construir?

En principi, el nombre de classes d'una taula de freqüències no està establert segons una llei, és el sentit comú el que prima a l'hora d'establir el nombre d'intervals. Així doncs, molt poques classes no revelen cap detall sobre les dades, i massa classes és tan confús, com la mateixa llista de dades originals.

1. Cal remarcar que aquesta representació no és l'única.

Malgrat tot, en la literatura matemàtica és possible trobar diverses regles per a calcular el nombre adequat d'intervals a partir del nombre de dades: la fórmula de Sturges o el mètode de l'arrel. Tot seguit es descriuen breument aquests dos darrers mètodes.

Mètode de l'arrel
Segons aquest mètode, el nombre de classes és igual a l'arrel quadrada del nombre de dades:

$$\text{Nombre de classes} \approx \sqrt{\text{nombre de dades}}$$

Fórmula de Sturges
Segons aquest mètode, el nombre d'intervals es calcula emprant la fórmula següent:

$$\text{Nombre de classes} \approx 1 + 3{,}32 \cdot \log(\text{nombre de dades})$$

Com ja s'ha esmentat anteriorment, no existeix un criteri per a escollir un mètode o l'altre. No obstant això, si el nombre de dades és molt gran, sol ser més convenient emprar la fórmula de Sturges. Tanmateix, si el nombre de dades és menut, és més convenient emprar el mètode de l'arrel.

El gràfic següent (figura 2) reflecteix el nombre d'intervals (eix y) segons el nombre de dades (eix x):

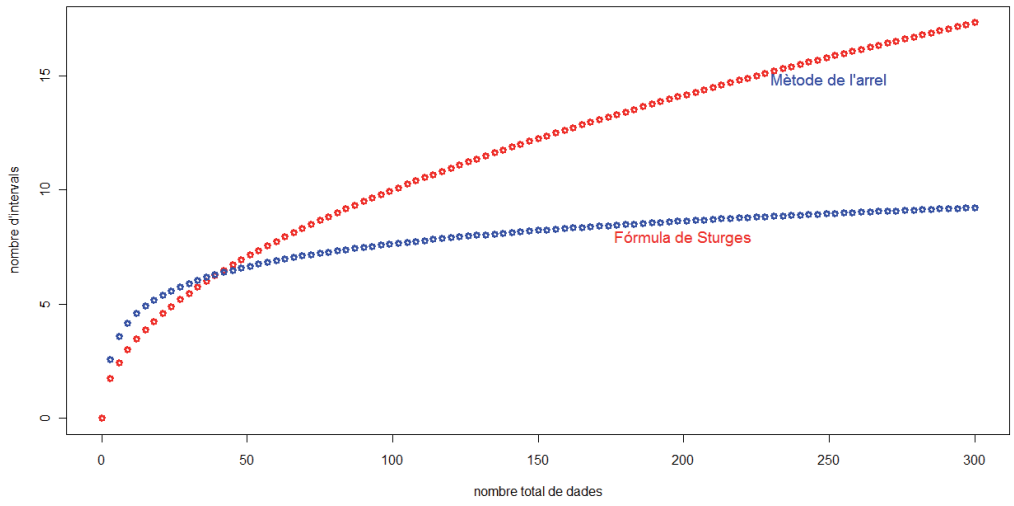


Figura 2

En els exemples que es consideraran al llarg del text s'emprarà el mètode de l'arrel, ja que el nombre de dades no serà massa gran.

Així, en l'exemple que estem considerant:

$$\text{Nombre de classes} \approx \sqrt{20} = 5.$$

El pas següent és calcular l'amplària dels intervals. Abans, però, és necessària una definició.

S'anomena *recorregut d'un conjunt de dades* la diferència entre el valor més gran i el més petit del conjunt. Es denota *Re*. A l'exemple el $Re = 15 - 5 = 10$.

En conseqüència, l'amplària $\approx \frac{\text{recorregut}}{\text{nombre de classes}}$. En l'exemple, l'amplària $\approx \frac{10}{5} = 2$.

Coneixent el nombre d'intervals i l'amplària es poden construir fàcilment tots els intervals (vegeu exemple 2). En finalitzar la construcció de tots els intervals és necessari comprovar que *totes* les dades pertanyen a un –i sols a un– interval. Si no és així, cal realitzar alguna modificació en l'amplària o en el nombre d'intervals.

En els exemples següents es posen de manifest totes aquestes qüestions:

Exemple 1

El govern desitja esbrinar si el nombre de fills per família ha descendit respecte de la dècada anterior. Per a això ha enquestat 50 famílies, i n'ha obtingut les dades següents respecte al nombre de fills:

2	4	2	3	1	2	4	2	3	0	2	2	2	3	2	6	2	3	2	2	3	2	3	3	4
3	3	4	5	2	0	3	2	1	2	3	2	2	3	1	4	2	3	2	4	3	3	2	2	1

Es demana:

a) Quina és la població objecte d'estudi.
b) Quina variable s'estudia.
c) Quin tipus de variable és.
d) Construir la taula de freqüències.
e) Quin és el nombre de famílies que tenen com a màxim 2 fills.
f) Quantes famílies tenen més d'1 fill, però com a màxim 3.
g) Quin percentatge de famílies té més de 3 fills.

Les respostes són:

a) La població objecte d'estudi és el conjunt de famílies d'un determinat país.
b) La variable que s'estudia és el nombre de fills per família.
c) El tipus de variable és quantitativa discreta, ja que el nombre de fills només pot prendre determinats valors enters (és impossible tenir mig fill...).
d) Seguint l'exemple, la taula queda:

x_i	n_i	N_i	f_i	F_i
0	2	2	0,04	0,04
1	4	6	0,08	0,12
2	21	27	0,42	0,54
3	15	42	0,30	0,84
4	6	48	0,12	0,96
5	1	49	0,02	0,98
6	1	50	0,02	1
	$N = 50$		1	

e) El nombre de famílies que tenen 2 o menys fills és: 2 + 4 + 21 = 27.
f) El nombre de famílies que tenen més d'1 fill però 3 com a màxim és: 21 + 15 = 36.
g) Les famílies que tenen més de 3 fills, són aquelles que tenen 4, 5 o 6 fills. Per tant, n'hi ha 6 + 1 + 1 = 8. El percentatge és la suma de les freqüències relatives dels valors multiplicat per 100:

$$(0,12 + 0,02 + 0,02) \cdot 100 = 0,16 \cdot 100 = 16\%$$

Exemple 2

En una certa ciutat s'obrirà un nou hotel. Abans de decidir el preu de les habitacions, el gerent investiga els preus per habitació de 40 hotels de la mateixa categoria d'aquesta ciutat. Les dades obtingudes en desenes d'euro són les següents:

3,9	4,7	3,7	5,6	4,3	4,9	5,0	6,1	5,1	4,5
5,3	3,9	4,3	5,0	6,0	4,7	5,1	4,2	4,4	5,8
3,3	4,3	4,1	5,8	4,4	4,8	6,1	4,3	5,3	4,5
4,0	5,4	3,9	4,7	3,3	4,5	4,7	4,2	4,5	4,8

Es demana:

a) Quina és la població objecte d'estudi.
b) Quina variable s'estudia.
c) Quin tipus de variable és.
d) Quin problema planteja la construcció de la taula de freqüències.
e) Quants hotels tenen un preu entre 3,25 i 3,75.
f) Quants hotels tenen un preu de 4,75 o superior.
g) Quin percentatge d'hotels té un cost menor que 4,25.

Les respostes són:

a) La població objecte d'estudi són els hotels de la ciutat.
b) La variable que s'estudia és el preu.
c) El tipus de variable és quantitativa contínua.
d) El problema que es planteja és que hi ha molts valors diferents. Per tant, és convenient agrupar la sèrie en intervals. Tenint en compte el que hem esmentat amb anterioritat:

El recorregut $Re = 6,1 - 3,3 = 2,8$; el nombre d'intervals és igual a l'arrel quadrada de 40, que és 6,32. Per tant, es prendran 6 intervals. Per a acabar, l'amplària dels intervals és el quocient entre el recorregut i el nombre d'intervals: 2,8/6 = 0,46.

Per a facilitar els càlculs aritmètics i la claredat de la taula de freqüències, es pren com a primer element 3,25 i com a amplària, 0,5. La taula queda així:

$[L_{i-1}, L_i)$	n_i	N_i	f_i	F_i
[3,25 – 3,75)	3	3	0,075	0,075
[3,75 – 4,25)	8	11	0,2	0,275
[4,25 – 4,75)	14	25	0,35	0,625
[4,75 – 5,25)	6	31	0,15	0,775
[5,25 – 5,75)	4	35	0,1	0,875
[5,75 – 6,25)	5	40	0,125	1
	$N = 40$			

e) 3.
f) 15.
g) 0,275 * 100 = 27,5%.

3.4. Gràfics estadístics

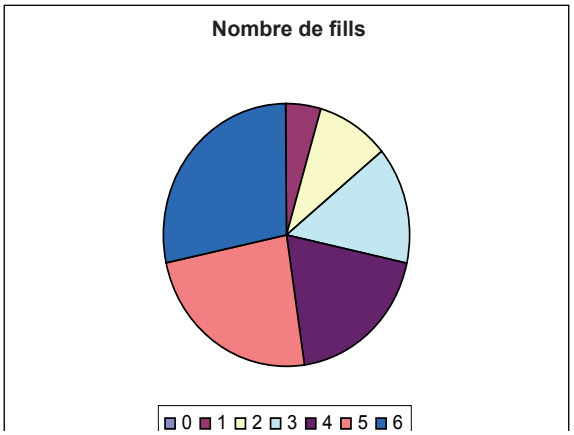
Els gràfics també són molt útils per a descriure els conjunts de dades. De fet, un gràfic estadístic permet formar-se una primera idea de la distribució de les dades tan sols amb una observació. No obstant això, cal anar amb compte, ja que en algunes ocasions els gràfics presenten tendències no atribuïbles al quefer matemàtic.

Es comentaran alguns tipus de gràfics, els més convencionals.

Diagrama de sectors o diagrama circular. És un cercle dividit en diferents sectors. L'àrea de cada sector és proporcional a la freqüència que es vulga representar, siga absoluta o relativa.

Per a calcular l'angle associat a cada freqüència s'hi aplica una simple proporció: l'angle associat a una freqüència absoluta n_i és igual a $f_i * 360º$. ($f_i = \frac{n_i}{n}$). Per a la freqüència absoluta acumulada es raonaria de la mateixa manera.

En l'exemple 1 de l'apartat anterior:

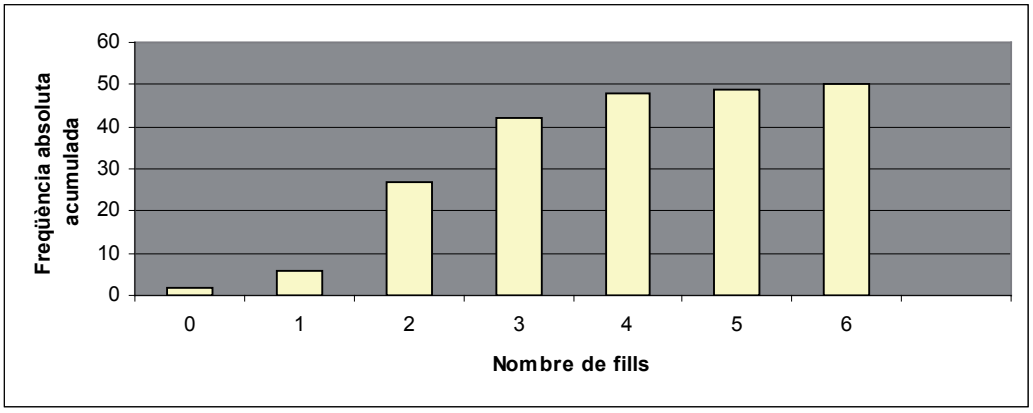


Nombre de fills	Freqüència absoluta	Angle
0	2	14,4
1	4	28,8
2	21	151,2
3	15	108
4	6	43,2
5	1	7,2
6	1	7,2

Aquest diagrama s'utilitza per a qualsevol tipus de variable estadística.

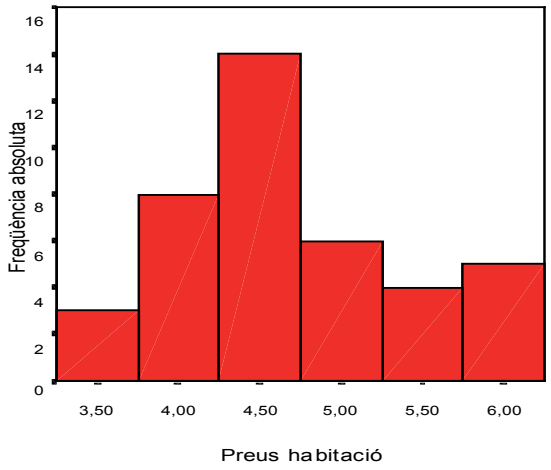
Diagrama de barres. S'utilitza per a representar les dades que no estan agrupades. Consisteix a col·locar sobre un eix horitzontal els diferents valors que pren la variable estadística, i sobre cadascun alçar un rectangle d'alçària igual a la freqüència (del tipus que s'estiga representant). Tots els rectangles han de tenir la mateixa amplària.

En l'exemple 1 tractat amb anterioritat:



Histogrames. S'utilitzen per a representar dades agrupades en intervals. Consisteix a col·locar sobre un eix horitzontal els diferents intervals. Sobre cadascun es construeix un rectangle de superfície igual a la freqüència que s'estiga representant. Així, les altures dels rectangles han de ser les densitats dels intervals.

No obstant això, en cas que totes les classes tinguen la mateixa amplària es pot utilitzar la freqüència com l'alçària del rectangle i no la densitat, ja que la forma del gràfic no varia. En al majoria dels casos és això mateix el que ocorre. En l'exemple 2:

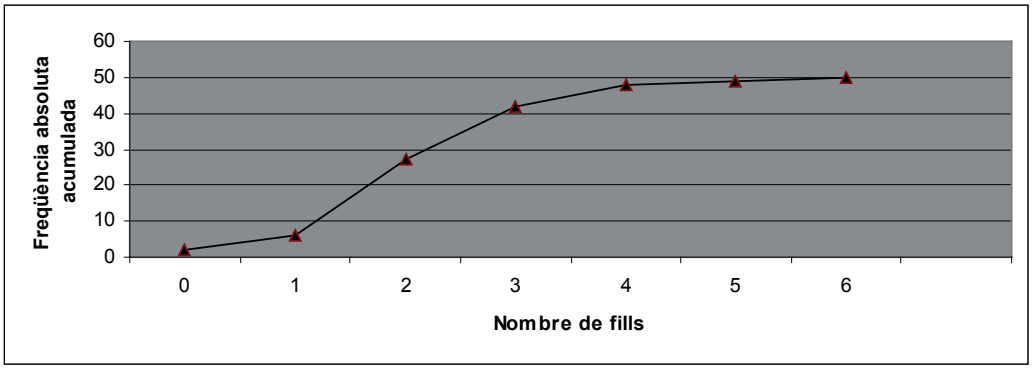


Preu	n_i
[3,25 – 3,75)	3
[3,75 – 4,25)	8
[4,25 – 4,75)	14
[4,75 – 5,25)	6
[5,25 – 5,75)	4
[5,75 – 6,25)	5

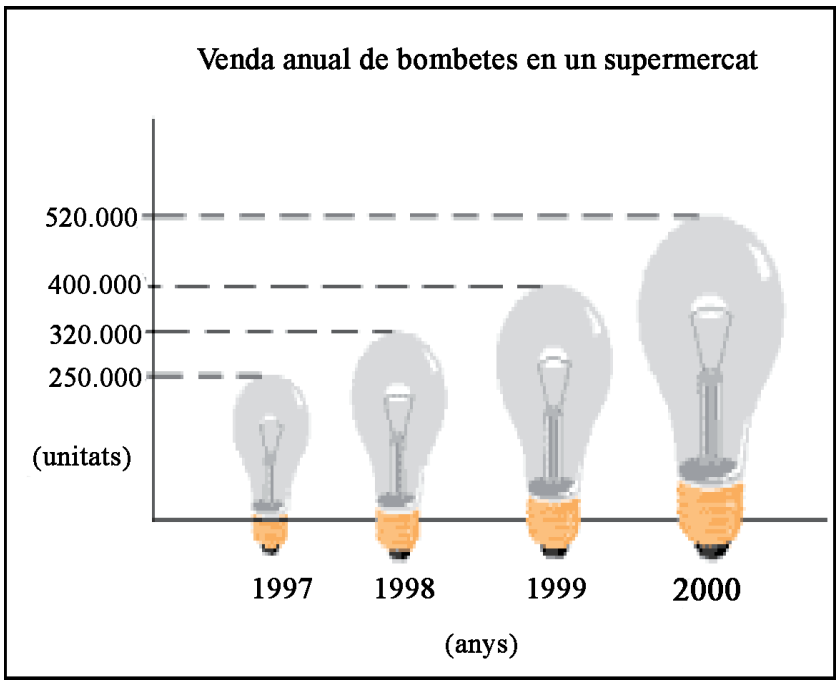
Cal notar que a l'eix horitzontal apareixen reflectides les marques de classe.

Polígon de freqüències. És menys utilitzat que els diagrames de barres i els histogrames, però pot substituir-los. Consisteix a unir, mitjançant línies poligonals, els extrems superiors de les barres, si es tracta de dades sense agrupar, o el punt mitjà de la base superior dels rectangles, si es tracta d'histogrames.

En l'exemple 1:



Pictograma. S'acostuma a utilitzar per a expressar un atribut. S'hi solen emprar icones que s'identifiquen amb la variable (per exemple, una bombeta) i la seua grandària és proporcional a la freqüència.



3.5. Problemes proposats

En aquest epígraf es plantejaran un conjunt de problemes per a la resolució dels quals és necessari conèixer la teoria desenvolupada al llarg de la unitat.

Exercici 1

El govern desitja saber si el nombre mitjà de fills per família ha descendit respecte de la dècada anterior. Per a això ha enquestat 50 famílies, respecte del nombre de fills, i n'ha obtingut les dades següents:

2	4	2	3	1	2	4	2	3	0	2	2	2	3	2	6	2	3	2	2	3	2	3	3	4
3	3	4	5	2	0	3	2	1	2	3	2	2	3	1	4	2	3	2	4	3	3	2	2	1

a) Construeix la taula de freqüències a partir d'aquestes dades.
b) Quantes famílies tenen exactament 3 fills?
c) Quin percentatge de famílies té exactament 3 fills?
d) Quin percentatge de les famílies de la mostra té més de 2 fills? I menys de 3?
e) Construeix el gràfic que consideres més adequat amb les freqüències no acumulades.
f) Construeix el gràfic que consideres més adequat amb les freqüències acumulades.

Exercici 2

En un hospital es desitja fer un estudi sobre els pesos dels xiquets de bolquers. Per a això, s'arrepleguen les dades de 40 xiquets i s'obté:

3,2	3,7	4,2	4,6	3,7	3,0	2,9	3,1	3,0	4,5
4,1	3,8	3,9	3,6	3,2	3,5	3,0	2,5	2,7	2,8
3,0	4,0	4,5	3,5	3,5	3,6	2,9	3,2	4,2	4,3
4,1	4,6	4,2	4,5	4,3	3,2	3,7	2,9	3,1	3,5

Es demana:

a) Construir-ne la taula de freqüències.
b) Si sabem que els xiquets que pesen menys de 3 quilos naixen prematurament, quin percentatge de xiquets prematurs han nascut entre aquests 40?
c) Normalment els xiquets que pesen més de 3 quilos i mig no necessiten estar a la incubadora. Pots dir quin percentatge de xiquets està en aquesta situació?
d) Representar gràficament la informació arreplegada.

Exercici 3

Abans de les últimes eleccions generals, una enquesta realitzada sobre la intenció de vot de col·lectiu de 45 persones, va donar els resultats següents:

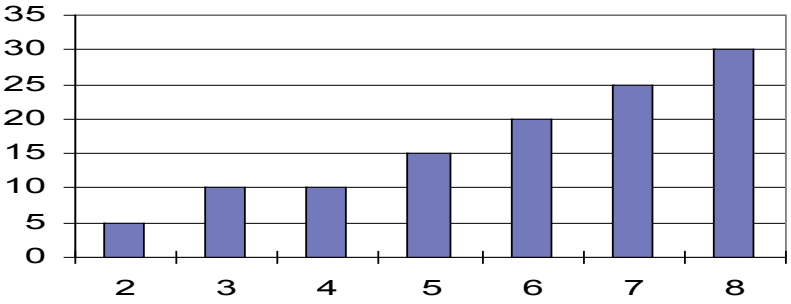
PP	PSOE	EU	PP	PSOE	UV	PP	UV	PSOE
EU	PP	EU	PP	UV	PP	PP	PSOE	UV
PSOE	PP	PSOE	UV	PP	UV	UV	PSOE	PP
EU	PP	PSOE	EU	PP	EU	UV	UV	PP
PSOE	UV	PP	PSOE	PP	EU	PP	EU	PP

Es demana:

a) Confeccionar una taula de freqüències que arreplegue aquesta informació i elaborar dos tipus de gràfics diferents a partir de la taula.
b) Quin percentatge de votants espera tenir cada formació política?

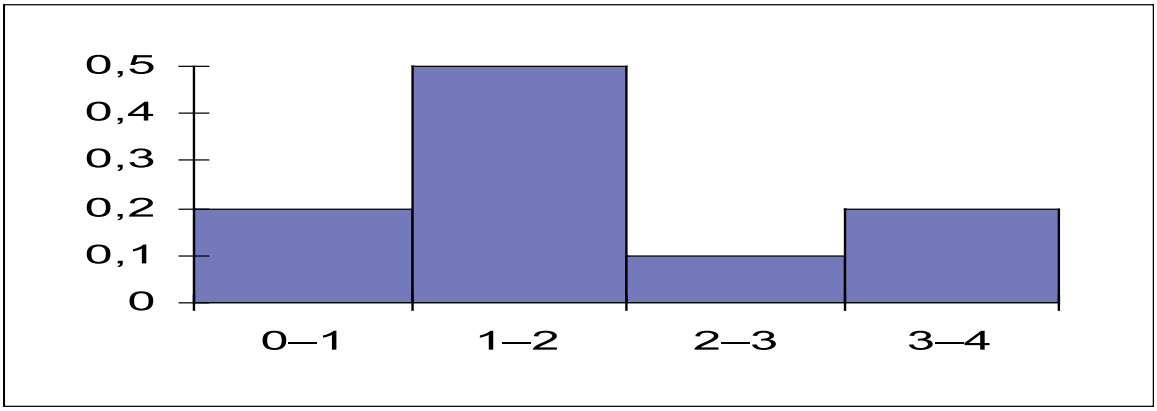
Exercici 4

Construeix la taula de freqüències a partir del següent gràfic de freqüències acumulades, sabent que tenim una variable discreta.



Exercici 5

Construeix la taula de freqüències relacionada amb el gràfic següent, on es resumeix dades arreplegades sobre 50 persones:



Exercici 6

Representa, mitjançant un histograma, els sous dels treballadors d'una empresa que apareixen en la taula:

Sous	n_i
[12,13)	20
[13,20)	25
[20,21)	10
[21,25)	10
[25,32)	8

Les dades estan donades en milers d'euros

TEMA 4

Distribució estadística d'una variable (II): mesures de posició i de dispersió

Objectius tema 4

- Conèixer els conceptes i saber realitzar els càlculs de les mesures de tendència central i de dispersió.
- Aplicar correctament les propietats de la mitjana aritmètica i de la variància.
- Aplicar el coeficient de variació de Pearson en aquelles situacions que ho requerisquen.
- Conèixer els principals estadístics que mesuren la forma dels gràfics.
- Saber interpretar i realitzar un diagrama de caixa.
- Saber calcular i interpretar l'índex de Gini, així com saber realitzar la corba de Lorenz per a mesurar l'equitat d'un repartiment.

4.1. Introducció

Tant les taules de freqüència com els gràfics tractats al capítol anterior permeten obtenir una gran quantitat d'informació del conjunt de dades. No obstant això, també existeixen altres mesures amb el mateix propòsit. Hi ha valors de la variable estadística al voltant dels quals les dades s'agrupen de la mateixa manera que les persones s'agrupen entorn d'un espectacle, o d'un esdeveniment festiu, etc. És a dir, les dades semblen amuntegades cap a un punt central anomenat *mesura central*. Així doncs, aquestes mesures estadístiques numèriques es poden considerar com una síntesi de tota la informació que les dades amaguen. Per exemple, no té el mateix significat que la nota mitjana d'un examen d'estadística siga un 3,5, que el fet que siga un 7,52. La informació que ens aporten aquests dos nombres dóna idea de la distribució de dades d'on provenen.

Tanmateix, no sempre aquestes mesures són del tot fiables. Únicament cal pensar en la situació següent: les notes obtingudes per 7 alumnes en un examen d'estadística han sigut 1, 1, 0, 0, 10, 10 i 10. La nota mitjana d'aquest grup és evidentment un 4,6. És obvi que en aquest cas la mitjana no és una mesura que reflectisca bé la informació que aporten les dades. Les dades no s'amunteguen entorn de 4,6 i, per tant, caldran altres mesures per a conèixer el grau de representativitat de les mesures de posició. Aquestes darreres mesures s'anomenen *mesures de dispersió* i indiquen fins a quin punt les dades s'escampen més o menys al voltant del punt central; per tant, són un reflex de la tendència de les observacions individuals a desviar-se de la mesura central.

Per altra part, també és interessant conèixer alguns valors de la variable estadística que, sense ser valors centrals de la distribució de dades, tenen importància per la informació que proporcionen. Aquests valors són, entre d'altres, els decils, els quartils, etc.

Al llarg d'aquesta unitat s'estudiaran les mesures de posició, tant de caràcter central com de no central, i les mesures de dispersió més usuals. Després farem referència a mesures que ens aportaran informació de la forma de l'histograma. Per a finalitzar el capítol, s'estudiaran algunes mesures de concentració, les quals poden aportar informació sobre el grau de repartiment del volum d'una magnitud, entre els diferents grups o sectors dels quals hem obtingut la quantitat total de la magnitud. Per exemple, informaran de si la nòmina que mensualment reparteix una empresa als treballadors està concentrada en algun grup de persones. De manera esquemàtica aquesta unitat tractarà:

Mesures de posició

Són coeficients que tracten de representar una determinada distribució, poden ser de dos tipus:

Centrals

Mitjanes
- Aritmètica
- Geomètrica
- Harmònica

Mediana
Moda

No centrals

Quartils
Decils
Percentils

Mesures de dispersió

Són complementàries de les de posició, en el sentit que assenyalen la dispersió del conjunt de totes les dades de la distribució, respecte de la mesura o mesures de localització adoptades.

Mesures de dispersió absoluta

Recorregut
Recorregut interquartílic
Desviació mitjana
Variància i desviació típica

Mesures de dispersió relativa

Coeficient de variació de Pearson

Mesures de forma

Ens donen informació de la forma de l'histograma, de la seua simetria i del grau de proximitat dels valors de la variable respecte de la mitjana.

Coeficient d'asimetria de Fisher
Coeficient de curtosi o apuntament
Diagrama de caixa

Mesures de concentració

Estudien el grau de concentració d'una magnitud, normalment econòmica, en determinats individus. En certa manera és un terme oposat a l'equitat en el repartiment.

Índex de Gini i corba de Lorenz

4.2. Mesures de posició

Es tractaran en primer lloc les mesures centrals i després les no centrals, seguint el guió establert a la introducció.

Cal remarcar que al llarg d'aquesta unitat, si no es comenta el contrari, es considerarà x la variable estudiada; el conjunt de nombres $\{x_1, x_2, ..., x_k\}$, els valors diferents de la variable; i n_i, la freqüència absoluta associada a cada X_i. Es considerarà també que el nombre total de dades és n.

4.2.1. Mesures de localització centrals. Mitjanes

Les mesures de localització centrals són els valors de la variable estadística al voltant dels quals s'acumulen un gran nombre de dades. Tot seguit es tractaran alguns d'aquests estadístics.

Mitjana aritmètica

És el valor que habitualment es pren com a representació de les dades. És la suma de tots els valors de la variable dividida entre el nombre total d'elements. Si les dades estan agrupades, es pren la marca de classe com a representant de l'interval i es realitzen tots els càlculs com si els valors de la variable foren les marques de classe.

Si es considera una variable estadística X que té k valors diferents, els quals es representen per x_i i les seues freqüències per n_i, llavors la mitjana aritmètica es calcula amb la fórmula següent:

$$\frac{n_1 x_1 + n_2 x_2 + ... + n_k x_k}{n}$$

Cal remarcar que quan les dades provenen d'una mostra, la mitjana aritmètica es denota per $\overline{X}$. Si, en canvi, es coneixen totes les observacions de la població, s'escriu la lletra grega μ.

Es considerarà que les dades provenen d'una mostra al llarg de la unitat. D'aquesta manera el símbol que s'utilitzarà és $\overline{X}$. Cal dir que aquest fet no constitueix una pèrdua de generalitat dels resultats.

Tenint en compte el que s'ha dit anteriorment, la notació amb el símbol sumatori ($\sum$) i que $n_1x_1 + n_2x_2 + \ldots\ldots + n_kx_k = \sum_{i=1}^{k} n_i \cdot x_i$, llavors una altra manera d'escriure la mitjana aritmètica és $\overline{X} = \frac{\sum_{i=1}^{k} n_i \cdot x_i}{n}$. A més a més, aprofitant que $f_i = \frac{n_i}{n}$ s'obté una altra expressió per a la mitjana: $\overline{X} = \sum_{i=1}^{k} f_i \cdot x_i$.

Nota

Encara que el càlcul manual de la mitjana aritmètica és senzill, cal dir que quasi totes les calculadores científiques admeten un procediment molt senzill en el paquet estadístic, que permet calcular-la amb molta facilitat. Evidentment, també els fulls de càlcul i els programes informàtics estadístics es poden emprar per a calcular aquest estadístic.

Exemple 1

La taula següent mostra el pes en quilograms de 10 xiquets d'una classe. Troba la mitjana aritmètica de la distribució de pesos.

x_i	n_i	$x_i n_i$
54	2	108
59	3	177
63	4	252
64	1	64
	$n = 10$	601

Llavors la mitjana aritmètica es calcula mitjaçant la fórmula:

$$\overline{X} = \frac{\sum_{i=1}^{k} n_i \cdot x_i}{n} = \frac{601}{10} = 60{,}1 \text{ kg}$$

Si la variable està agrupada en intervals, el concepte no canvia. En aquest cas, s'assignen les freqüències a les marques de classe i es procedeix de la mateixa manera que en el cas de les no agrupades.

Cal notar que en el futur es considerarà indistintament $c_i = x_i$.

Exemple 2

La taula següent mostra el pes de 10 xiquets d'una classe però aquesta vegada les dades estan agrupades en intervals.

$[L_{i-1}, L_i)$	$x_i = c_i$	n_i	$c_i \cdot n_i$
[30, 40)	35	3	105
[40, 50)	45	2	90
[50, 60)	55	5	275
		$n = 10$	470

En aquest cas la mitjana aritmètica és:

$$\overline{X} = \frac{\sum_{i=1}^{k} n_i \cdot c_i}{n} = \frac{470}{10} = 47 \text{ kg}$$

Nota

La mitjana aritmètica s'anomena també *centre de gravetat de la distribució*, ja que n'és el punt d'equilibri. Gràficament correspon a la descripció següent: en una variable sense agrupar, considerem els valors que pren la variable estadística situats sobre una vareta ideal sense pes, de manera que sobre cada valor de la variable s'ubique un pes igual a la freqüència. El punt de la vareta que l'equilibra és la mitjana aritmètica (figura 1). D'una manera similar es raona amb les dades agrupades.

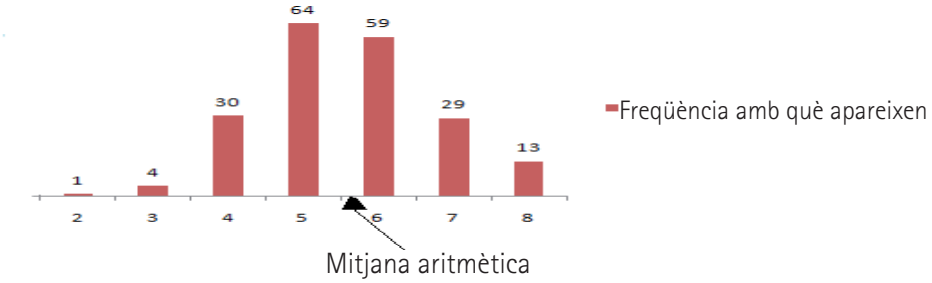


Figura 1

El concepte *centre de gravetat* s'observa en el fet que, si la variable estadística presenta valors extrems, llavors la mitjana aritmètica, per a intentar equilibrar, s'hi desplaça dels del centre del conjunt de dades.

En aquest gràfic (figura 2) existeix un valor de la variable estadística amb molta freqüència i la mitjana s'hi deplaça.

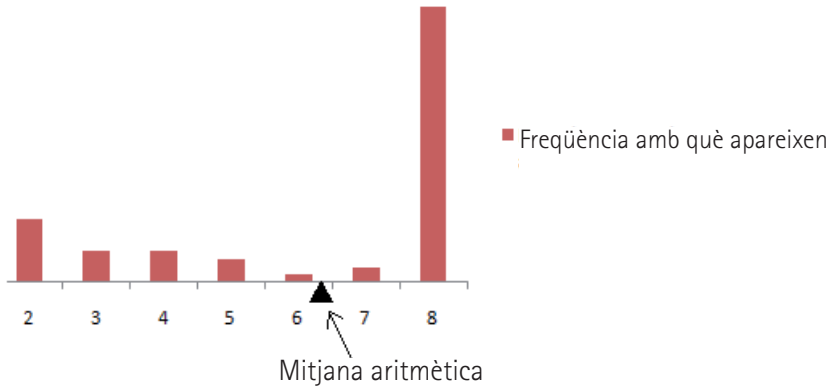


Figura 2

Per contra, en aquest tercer gràfic (figura 3) hi ha un valor de la variable molt diferent de la resta, i la mitjana també se'n veu afectada, ja que s'hi desplaça. En ambdós casos s'observa que la mitjana aritmètica no és tan representativa de les dades en conjunt.

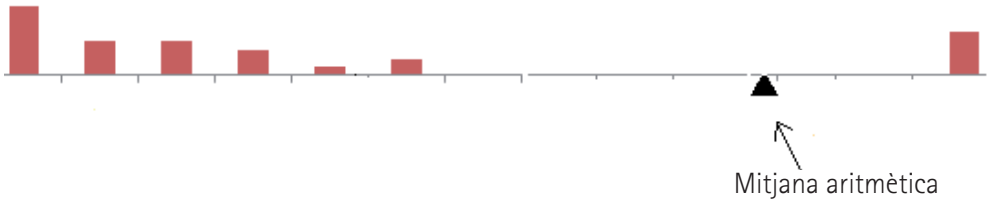


Figura 3

Propietats de la mitjana aritmètica

La mitjana aritmètica compleix les propietats següents:

Propietat 1
La suma de les desviacions dels valors de la variable respecte a la mitjana aritmètica és 0.

Si s'opera sobre l'expressió $\sum_{i=1}^{k}(x_i - \overline{X})n_i$, es té que:

$$\sum_{i=1}^{k}(x_i - \overline{X})n_i = \sum_{i=1}^{k}x_i n_i - \sum_{i=1}^{k}\overline{X}n_i = \sum_{i=1}^{k}x_i n_i - \overline{X}\sum_{i=1}^{k}n_i = n\frac{\sum_{i=1}^{k}x_i n_i}{n} - \overline{X}n = n\overline{X} - \overline{X}n = 0,$$

fet que demostra la propietat.

És a dir:

La desviació d'un valor de la variable estadística respecte de la mitjana aritmètica és la diferència $x_i - \overline{X}$. Com que la freqüència del valor x_i és n_i, llavors la suma de les desviacions de tots els valors iguals a x_i és $n_i(x_i - \overline{X})$. A més, la propietat diu que en sumar les desviacions de totes les dades el resultat serà 0. Llavors s'obté: $(x_1 - \overline{X})n_1 + (x_2 - \overline{X})n_2 + ... + (x_k - \overline{X})n_k = \sum_{i=1}^{n}(x_i - \overline{X})n_i = 0$.

Ho comprovarem amb un simple exemple. Se suposa la variable estadística X = nombre de fills per família en un barri de Borriana. Les dades obtingudes han estat 3, 3, 2, 2, 3, 2, 3, 2, 1, 4, i la mitjana aritmètica n'és 2,5 fills per família.

Tenint present que els valors que pren la variable són 1, 2, 3 i 4 amb freqüències absolutes 1, 4, 4 i 1, respectivament, la suma de les desviacions serà:

$(x_1 - \overline{X})n_1 + (x_2 - \overline{X})n_2 + ... + (x_k - \overline{X})n_k = (1 - 2{,}5) \cdot 1 + (2 - 2{,}5) \cdot 2 +$
$+ (3 - 2{,}5) \cdot 2 + (3 - 2{,}5) \cdot 2 + (4 - 2{,}5) \cdot 1 = -1{,}5 - 2{,}5 + 2{,}5 + 1{,}5 = 0.$

Propietat 2
Si a tots els valors de la variable se suma una mateixa constant, la mitjana aritmètica queda augmentada amb la dita constant.

Se suposa que s'està treballant amb una variable X de la qual es coneix la mitjana aritmètica.

Si ara es defineix una altra variable Y, a partir de l'anterior, sumant a cada valor de la variable X una constant C, és a dir, els valors de la variable Y són els generats per la variable X segons la relació següent: $y_i = x_i + C$, la mitjana aritmètica d'aquesta segona variable es pot calcular afegint la constant C a la mitjana aritmètica de la variable X.

En efecte, amb les condicions del paràgraf anterior, $y_i = x_i + C$, es té que:

$$\overline{Y} = \frac{\sum_{I=1}^{k} y_i n_i}{n} = \frac{\sum_{I=1}^{k}(x_i + C)n_i}{n} = \frac{\sum_{I=1}^{k} x_i n_i + \sum_{I=1}^{k} Cn_i}{n} = \frac{\sum_{I=1}^{k} x_i n_i}{n} + \frac{\sum_{I=1}^{k} Cn_i}{n} = \frac{\sum_{I=1}^{k} x_i n_i}{n} + \frac{Cn}{n} = \overline{X} + C.$$

Ja que $\frac{\sum_{I=1}^{k} x_i n_i}{n} = \overline{X}$ i $\sum_{i=1}^{k} n_i = n$. Llavors, en substituir es té que $\overline{Y} = \overline{X} + C$.

És a dir:

El fet que una variable es genere a partir d'una altra sumant-li una constant $Y = X + C$ significa el que es mostra tot seguit:

Valors de la variable inicial	*Valors de la variable modificada*
x_1	$y_1 = x_1 + c$
x_2	$y_2 = x_2 + c$
....	.....
x_k	$y_k = x_k + c$

La propietat afirma que, si això és així, llavors $\overline{Y} = \overline{X} + C$.

Exemple 3

Els sous de 20 treballadors d'una empresa l'any 2006 es mostren segons la taula següent:

***X* = sous (€)**	**Treballadors**
17.500	12
22.500	6
27.500	1
32.500	1

La mitjana aritmètica dels sous és 20.250 €.

Si l'any 2007 l'empresa decideix apujar el sou de cada treballador 1.000 €, llavors s'obté una nova variable Y (Y = sou dels treballadors l'any 2007) a partir de l'anterior. L'equació que relaciona totes dues variables és:

$$\text{Sous any 2007} \qquad Y = X + 1.000$$

A partir de la relació entre les variables s'obté la taula següent. I conseqüentment, es pot calcular el sou mitjà de l'any 2007, $\overline{Y} = 21.250$ €.

***Y* = sous (€)**	**Treballadors**
18.500	12
23.500	6
28.500	1
33.500	1

Però, si únicament interessara el sou mitjà $\overline{Y}$, no caldria construir aquesta taula de freqüències, simplement s'aplicaria la propietat anterior i ens estalviaríem els càlculs:

$$\overline{Y} = \overline{X} + C = 20.250 + 1.000 = 21.250 \text{ €}$$

Ja que $C = 1.000$ i $\overline{X} = 20.250$.

Propietat 3
Si tots els valors de la variable es multipliquen per una mateixa constant, la mitjana aritmètica queda multiplicada per la dita constant.

Se suposa que s'està treballant amb una variable X de la qual es coneix la mitjana aritmètica.

Si ara es defineix una altra variable Y a partir de l'anterior multiplicant cada valor de la variable X per una constant a, és a dir, els valors de la variable Y són els generats per la variable X segons la relació següent: $y_i = ax_i$. La mitjana aritmètica d'aquesta segona variable es pot calcular multiplicant la mitjana aritmètica de la variable X per a.

En efecte, amb les condicions del paràgraf anterior, $y_i = ax_i$, es té que:

$$\overline{Y} = \frac{\sum_{i=1}^{k} y_i n_i}{n} = \frac{\sum_{i=1}^{k} ax_i n_i}{n} = \frac{a\sum_{i=1}^{k} x_i n_i}{n} = a\frac{\sum_{i=1}^{k} x_i n_i}{n} = a\overline{X}$$

És a dir:

El fet que una variable es genere a partir d'una altra multiplicant-la per una constant, $Y = a \cdot X$, significa el que es mostra tot seguit:

Valors de la variable inicial	*Valors de la variable modificada*
x_1	$y_1 = a \cdot x_1$
x_2	$y_2 = a \cdot x_2$
....	.....
x_k	$y_k = a \cdot x_k$

La propietat afirma que, si això és així, llavors $\overline{Y} = a \cdot \overline{X}$.

Exemple 4

Es consideren les mateixes dades que a l'exemple 3 anterior, és a dir, X = sous de 20 treballadors l'any 2006.

Se suposa que l'empresari decideix apujar el sou de cada treballador per a l'any 2007 un 2%. D'aquesta manera sorgeix una altra variable Y (Y = sou dels treballadors l'any 2007) en funció de la variable X. L'equació que relaciona totes dues variables és $Y = 1{,}02 \cdot X$ (apujar una quantitat un 2% és equivalent a multiplicar-la per 1,02).

Per a calcular la mitjana aritmètica de la variable Y (sou mitjà l'any 2007) no cal construir-ne la taula de freqüències, tan sols aplicar la propietat. Així:

$\overline{Y} = 1{,}02 \cdot 20.250 = 20.655$ €.

Es deixa com a exercici construir la taula de freqüències de la variable Y i comprovar que el càlcul de la mitjana a partir de la taula coincideix amb el que es busca mitjançant la propietat.

Propietat 4
Si una variable Y *és transformació lineal d'una altra variable* X *(* Y = a · X + b*; a i* b *nombres reals), la mitjana aritmètica de* Y *segueix la mateixa transformació lineal respecte a la mitjana aritmètica de* X*. És a dir:* $\overline{Y} = a \cdot \overline{X} + b$.

Se suposa que s'està treballant amb una variable X de la qual es coneix la mitjana aritmètica. En aquest cas la relació entre totes dues variables és $Y = a \cdot X + b$ i, per tant, cada dada x_i genera la dada y_i corresponent segons l'equació $y_i = a\,x_i + b$.

En efecte, amb les condicions del paràgraf anterior es té que:

$$\overline{Y} = \frac{\sum_{I=1}^{k} y_i n_i}{n} = \frac{\sum_{I=1}^{k} (ax_i + b) n_i}{n} = \frac{\sum_{I=1}^{k} ax_i n_i + \sum_{I=1}^{k} bn_i}{n} = \frac{a\sum_{I=1}^{k} x_i n_i}{n} + \frac{b\sum_{I=1}^{k} n_i}{n} = a\frac{\sum_{I=1}^{k} x_i n_i}{n} + \frac{bn}{n} = \overline{aX} + b.$$

És a dir:

El fet que una variable es genere a partir d'una altra multiplicant-la per un nombre i sumant-li una constant ($Y = a \cdot X + b$) significa el que es mostra tot seguit:

Valors de la variable inicial	*Valors de la variable modificada*
x_1	$y_1 = a \cdot x_1 + b$
x_2	$y_2 = a \cdot x_2 + b$
....	.....
x_k	$y_k = a \cdot x_k + b$

La propietat afirma que, si això és així, llavors $\overline{Y} = a \cdot \overline{X} + b$.

Exemple 5

Es consideren les mateixes dades que a l'exemple 4, és a dir, X = sous de 20 treballadors l'any 2006. Se suposa ara que l'empresari decideix apujar el sou de cada treballador per a l'any 2007 un 2% i, a més, 1.000 € en concepte de prima. D'aquesta manera sorgeix una altra variable $Y = 1{,}02\,X + 1.000$. Per a calcular la mitjana aritmètica de la variable Y (sou mitjà de l'any 2007) no cal construir la taula de freqüències, tan sols aplicar la propietat. Així:

$$\overline{Y} = 1{,}02 \cdot 20.250 + 1.000 = 21.655\ €.$$

Es deixa com a exercici construir la taula de freqüències de la variable Y i comprovar que el càlcul de la mitjana, a partir de la taula, coincideix amb el que es busca mitjançant la propietat.

Propietat 5

Si en un conjunt de valors es poden obtenir dos o més subconjunts disjunts que suposen una partició del conjunt total de valors, la mitjana aritmètica del conjunt es relaciona amb la mitjana aritmètica de cada un dels subconjunts disjunts de la forma següent: $\overline{\mathrm{X}} = \frac{\sum \overline{\mathrm{X}}_{\mathrm{i}} \cdot \mathrm{N}_{\mathrm{i}}}{\mathrm{n}}$ *(on* $\overline{\mathrm{X}}_{\mathrm{i}}$ *és la mitjana de cada subconjunt i* N_{i}*, el nombre d'elements de cada subconjunt).*

Seguidament se n'esbossa la demostració. S'hi suposen únicament dos subconjunts disjunts. Si n'hi haguera més, es demostraria de manera semblant.

Es considera la distribució de dades $x_1, x_2, \ldots x_{N_1}, x_{N_1+1}, x_{N_1+2}, \ldots, x_k \ldots$, on hi ha dos subconjunts de N_1 i $k - N_1 = N_2$ elements cada un.

La mitjana aritmètica de la distribució és $\overline{X} = \frac{\sum_{i=1}^{k} x_i n_i}{n}$. Desglossant el sumatori en dos (un per a cada subconjunt), l'expressió de la mitjana queda:

$$\overline{X} = \frac{\sum_{i=1}^{k} x_i n_i}{n} = \frac{\sum_{i=1}^{N_1} x_i n_i + \sum_{j=N_1+1}^{k} x_j n_j}{n} = \frac{\sum_{i=1}^{N_1} x_i n_i}{n} + \frac{\sum_{j=N_1+1}^{k} x_j n_j}{n} = (*)$$

Així, si es multipliquen el numerador i el denominador de cada una de les fraccions per una mateixa quantitat, el resultat no varia; per tant, multiplicant la primera per N_1 que és el nombre d'elements del primer subconjunt, i la segona per N_2, que és el nombre d'elements del segon subconjunt, l'expressió quedarà:

$$(*) = \frac{\frac{N_1 \sum_{i=1}^{N_1} x_i n_i}{N_1}}{n} + \frac{\frac{N_2 \sum_{j=N_1+1}^{k} x_j n_j}{N_2}}{n} = \frac{N_1 \frac{\sum_{i=1}^{N_1} x_i n_i}{N_1}}{n} + \frac{N_2 \frac{\sum_{j=N_1+1}^{k} x_j n_j}{N_2}}{n} = (**)$$

i com que $\overline{X_1} = \frac{\sum_{i=1}^{N_1} x_i n_i}{N_1}$ i $\overline{X_2} = \frac{\sum_{i=N_1+1}^{k} x_i n_i}{N_2}$ són les mitjanes del primer i el segon subconjunt, l'expressió la podem expressar de la manera següent:

$(**) = \frac{N_1 \overline{X_1}}{n} + \frac{N_2 \overline{X_2}}{n} = \frac{N_1 \overline{X_1} + N_2 \overline{X_2}}{n}$, que és el que es volia demostrar.

Nota

Cal mencionar que si en els subconjunts hi ha dades iguals, la propietat es continua complint. Únicament cal prestar atenció a la demostració de la propietat. L'exemple següent aclarirà aquest fet.

Exemple 6

Es considera la variable X = nombre de cotxes per família. Tres entrevistadors han obtingut les dades següents:

- Entrevistador A: {1, 2, 4, 3, 4, 5, 6, 7} $\quad \overline{X_A} = 4$
- Entrevistador B: {5, 6, 7, 3, 4, 5} $\quad \overline{X_B} = 5$
- Entrevistador C: {5, 6, 7} $\quad \overline{X_C} = 6$

Com calcular la mitjana conjunta, és a dir, la mitjana de les dades {1, 2, 4, 3, 4, 5, 6, 7, 5, 6, 7, 3, 4, 5, 5, 6, 7}?

$$\overline{X} = \frac{8 \cdot \overline{X_A} + 6 \cdot \overline{X_B} + 3 \cdot \overline{X_C}}{17} = \frac{8 \cdot 4 + 6 \cdot 5 + 3 \cdot 6}{17} = \frac{80}{17} = 4{,}71$$

Aplicant-hi la propietat, es pot comprovar fàcilment que la mitjana aritmètica del conjunt de dades {1, 2, 4, 3, 4, 5, 6, 7, 5, 6, 7, 3, 4, 5, 5, 6, 7}coincideix amb 4,71.

Mitjana aritmètica ponderada

De vegades, no tots els valors de la variable tenen el mateix pes. És a dir, cadascun dels valors que pren la variable té assignat un nombre que n'indica la importància, i el qual és independent de la mateixa freqüència absoluta.

El càlcul de la mitjana aritmètica ponderada en aquests casos segueix l'expressió, següent, on w_i és el pes associat a cada valor de la variable x_i:

$$\overline{X}_W = \frac{\sum_{i=1}^{k} x_i w_i n_i}{\sum_{i=1}^{k} w_i n_i}$$

Exemple 7

El personal d'una empresa ha estat avaluat recentment per una companyia especialitzada. L'empresa està dividida en quatre seccions. La taula següent mostra el nombre de treballadors de cada secció i la seua puntuació mitjana, respectivament. Es vol conèixer la valoració mitjana dels treballadors de l'empresa.

Secció	Nombre de treballadors	Valoració
1	10	5
2	20	8
3	30	7
4	30	7

Per a obtenir la valoració mitjana de tots els treballadors de l'empresa cal aplicar la mitjana aritmètica ponderada de les puntuacions de cada secció, ja que cada secció té una quantitat de treballadors diferent i, per tant, la puntuació de cada secció no té la mateixa importància o pes. Aquest pes està donat, evidentment, pel nombre de treballadors.

Aleshores, per a calcular la mitjana aritmètica ponderada podem construir la taula següent:

x_i	w_i	n_i	$x_i \cdot w_i \cdot n_i$	$w_i \cdot n_i$
5	10	1	50	10
8	20	1	160	20
7	30	2	420	60
			630	90

En conseqüència, la mitjana aritmètica ponderada és:

$$\overline{X}_W = \frac{\sum_{i=1}^{k} x_i w_i n_i}{\sum_{i=1}^{k} w_i n_i} = \frac{630}{90} = 7$$

Per tant, la valoració mitjana dels treballadors de l'empresa és 7.

Exemple 8

Un estudiant realitza tres exàmens de complexitat creixent (el criteri escollit per a determinar la complexitat de l'examen és el límit de temps), i obté els resultats següents: 5, 8 i 7.

Per a realitzar el primer examen se li concedeix una hora i mitja; per al segon, una hora; i per al tercer, mitja hora.

És evident que tots els exàmens no tenen el mateix valor, ja que la dificultat no és la mateixa. Per això cal fer-ne la ponderació. Es pot associar el nombre 1 a l'examen de menor dificultat (el primer), el 2 a l'examen de dificultat mitjana (el segon) i el 3 a l'examen més complex.

Cal notar que aquesta associació és coherent perquè en la ponderació s'associa una unitat de pes a un temps d'una hora i mitja.

x_i	n_i	w_i	$x_i \cdot w_i \cdot n_i$	$w_i \cdot n_i$
5	1	1	5	1
8	1	2	16	2
7	1	3	21	3
			42	6

$$\overline{X}_W = \frac{\sum_{i=1}^{k} x_i w_i n_i}{\sum_{i=1}^{k} w_i n_i} = \frac{42}{6} = 7$$

Si es calculara la mitjana aritmètica sense tenir en compte els pesos, totes les notes valdrien el mateix, un terç de la nota final. En l'exemple no és així. Per això es fa necessari calcular la mitjana aritmètica ponderada.

Mitjana geomètrica

Pot utilitzar-se per a mostrar canvis percentuals en una sèrie de nombres positius. Per tant, té una àmplia aplicació en els negocis i en l'economia. La mitjana geomètrica proporciona una mesura precisa d'un canvi percentual mitjà en una sèrie de nombres. Es representa per *G* i el seu càlcul –segons la notació habitual– segueix l'expressió següent:

x_i	n_i
x_1	n_1
x_2	n_2
x_3	n_3
.....	.....

$$G = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots\ldots x_k^{n_k}}$$

Emprant la notació potencial, també es pot representar com:

$$G = (x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots\ldots x_k^{n_k})^{\frac{1}{n}}$$

Exemple 9

La direcció d'una empresa desitja conèixer la taxa de creixement mitjana dels ingressos d'una secció de l'empresa al llarg dels anys 2003, 2004 i 2005. Si aquesta taxa és inferior a un 10% l'empresa prescindirà de la secció en breu. Els ingressos de la secció per anys es mostren en la taula:

Any	Ingressos	Índex de variació
2002	120.000 €	
2003	130.000 €	130.000/120.000 = 1,083
2004	143.000 €	143.000/130.000 = 1,1
2005	159.000 €	159.000/143.000 = 1,11

El primer que cal conèixer és el percentatge que representen els ingressos de cada any respecte dels obtinguts l'any anterior (índex de variació). Això és el que s'ha calculat a la tercera columna de la taula. Així, per exemple, els ingressos de l'any 2003 representen l'1,083% dels obtinguts l'any anterior. És a dir, han augmentat un 8,3% respecte de l'any 2002.

Per a calcular la taxa d'interès mitjana, cal trobar un percentatge i de manera que, si tots els anys hagueren augmentat els ingressos, aquest percentatge i, en finalitzar l'any 2005 els ingressos de la secció serien 159.000 €. És a dir:

Any	Índex de variació real	Capital acumulat (milers)	Índex de variació mitjà	Capital acumulat (milers)
2003	1,083	$120 \cdot 1{,}083$	$(1 + i)$	$120 \cdot (1 + i)$
2004	1,1	$120 \cdot 1{,}083 \cdot 1{,}1$	$(1 + i)$	$120 \cdot (1 + i)^2$
2005	1,11	$120 \cdot 1{,}083 \cdot 1{,}1 \cdot 1{,}11$	$(1 + i)$	$120 \cdot (1 + i)^3$

Com que en finalitzar l'any 2005 els ingressos han de ser els mateixos, es compleix la igualtat següent:

$$120 \cdot (1{,}083) \cdot (1{,}1) \cdot (1{,}11) = 120 \cdot (1 + i)^3$$

D'on s'obté que: $(1{,}083) \cdot (1{,}1) \cdot (1{,}11) = (1 + i)^3 \rightarrow (1 + i) = \sqrt[3]{(1{,}083) \cdot (1{,}1) \cdot (1{,}1)} = 1{,}097609$

És a dir, $(1 + i)$ es pot calcular com la mitjana geomètrica dels percentatges que representen els ingressos de cada any respecte dels ingressos dels anys anteriors, és a dir, la mitjana geomètrica dels índexs de variació.

Com que, per altra part, allò que es busca és la taxa de creixement mitjana, i, és necessari restar 1 a la mitjana geomètrica calculada i expressar-la en tant per cent. Aleshores $i = 9{,}7609\%$.

També se n'hauria pogut construir la taula de freqüències. Cal remarcar, però, que és necessari identificar la variable x_i.

En l'exemple, x_i = percentatge que representen els ingressos de cada any respecte dels obtinguts l'any anterior, és a dir, els índexs de variació.

La taula queda:

x_i	n_i
1,083	1
1,1	1
1,11	1

Aplicant-hi la fórmula $G = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot .....x_k^{n_k}}$: $G = \sqrt[3]{(1{,}083) \cdot (1{,}1) \cdot (1{,}1)} = 1{,}097609$.

I, per tant, la taxa d'interès mitjà és 0,097609, com s'havia deduït anteriorment.

Una altra manera de calcular la mitjana geomètrica és mitjançant els logaritmes.

$$G = \text{inversa del logaritme de} \left(\frac{\sum_{i=1}^{k} n_i \cdot \log x_i}{n} \right)$$

Així doncs, el logaritme de la mitjana geomètrica és la mitjana aritmètica dels logaritmes dels valors de la variable. Per altra part, les propietats dels logaritmes permeten demostrar l'equivalència de les dues expressions de *G*.

L'exemple clarificarà el procediment de càlcul de la mitjana geomètrica.

Exemple 10

Se suposa que la variable x_i representa l'índex de variació del valor d'uns productes d'una empresa en la borsa al llarg de 22 anys. Es desitja saber quina ha estat la taxa de variació mitjana del valor dels productes de l'empresa en els 22 anys.

x_i	n_i
1,083	10
1,120	5
1,125	4
1,140	3
	$n = 22$

S'han de calcular, en primer lloc, els logaritmes per a obtenir:

$$\log G = \frac{\sum_{i=1}^{k} n_i \cdot \log x_i}{n}.$$

Per tant, serà convenient ampliar la taula:

x_i	n_i	$\log x_i$	$n_i \cdot \log x_i$
1,083	10	log 1,083 = 0,035	0,35
1,120	5	log 1,120 = 0,049	0,245
1,125	4	log 1,125 = 0,051	0,204
1,140	3	log 1,140 = 0,057	0,171
	$N = 22$		0,97

$$G = \text{inversa logaritme}\left(\frac{\sum_{i=1}^{k} n_i \cdot \log x_i}{n}\right) = \text{inv. logaritme}\left(\frac{0,97}{22}\right) = \text{inv. log}\,(0,044) = 1,1069.$$

És a dir, la taxa de variació mitjana dels productes ha estat un 10,69%.

Nota

Cal recordar que l'operació inversa del logaritme és l'exponencial. Si en l'exemple utilitzem logaritme de base decimal, el que cal escriure serà $G_k = 10^{0,004} = 1,1069$.

Si utlitzàrem logaritmes neperians, l'exponencial seria $G = e^{\frac{\sum_{i=1} n_i \cdot \log x_i}{n}}$.

Mitjana harmònica

Es representa per H i és la inversa de la mitjana aritmètica de les inverses dels valors de la variable, respon a l'expressió següent:

$$H = \frac{n}{\sum \frac{n_i}{x_i}} = \frac{n}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + ... + \frac{n_k}{x_k}}$$

n = nombre total de dades
x_i = valors diferents que pren la variable
n_i = freqüencia absoluta de x_i

S'utilitza per a calcular el valor mitjà de magnituds expressades en termes relatius, com ara velocitats, temps, rendiment, tipus de canvi monetari, etc. La principal contrarietat és que quan algun valor de la variable és 0 o pròxim a 0 no es pot calcular.

En moltes ocasions, no és necessari aplicar la fórmula anterior. Únicament cal tenir present el concepte *mitjana aritmètica*.

Exemple 11

En un magatzem hi ha tres màquines que tenen les produccions següents:

- màquina 1 produeix 4 peces cada hora.
- màquina 2 produeix 20 peces cada hora.
- màquina 3 produeix 16 peces cada hora.

Si de la màquina 1 es van obtenir 400 peces, de la màquina 2 se'n van obtenir 500 i de la 3, 800, i les màquines no poden estar en funcionament simultàniament, quina és la mitjana de peces que va produir el magatzem per hora?

La mitjana que demana l'exemple es pot calcular com el quocient entre el nombre de peces totals que es van produir al magatzem dividit pel nombre d'hores que va costar produir-les. Així:

$$\text{mitjana} = \frac{\text{nombre total de peces}}{\text{nombre total d'hores per a produir-les}}.$$

El nombre total de peces és quasi directe: $400 + 500 + 800 = 1.700$ peces.

Tanmateix, el nombre d'hores emprades no és tan directe, ja que cada màquina no té la mateixa producció per hora.

Així, la màquina 1 utilitza $\frac{400}{4} = 100$ hores per a fer treball, la màquina 2 en necessita $\frac{500}{20} = 25$ i la màquina 3 hi inverteix $\frac{800}{16} = 50$ hores.

Així doncs, s'obté:

$$\text{mitjana} = \frac{\text{nombre total de peces}}{\text{nombre total d'hores que va costar produir-les}} =$$

$$= \frac{1.700}{\frac{400}{4} + \frac{500}{20} + \frac{800}{16}} = \frac{1.700}{100 + 25 + 50} = \frac{1.700}{175} = 9{,}71 \text{ peces per hora.}$$

Cal destacar que allò que s'ha calculat és efectivament la mitjana harmònica. Així, si es col·loquen les dades en una taula queda:

x_i	n_i
4	400
20	500
16	800
	1.700

Aplicant-hi la fórmula:

$$H = \frac{n}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \ldots + \frac{n_k}{x_k}} = \frac{1.700}{\frac{400}{4} + \frac{500}{20} + \frac{800}{16}} = 9{,}71 \text{ peces.}$$

Entre la mitjana aritmètica, la mitjana geomètrica i la mitjana harmònica es dóna sempre la relació següent: $H \leq G \leq \overline{X}$.

4.2.2. Mesures de localització centrals. Mediana

La mediana és el valor de la variable que divideix les observacions en dos grups d'igual nombre d'elements de manera que en el primer grup totes les dades siguen inferiors o iguals a la mediana, i en l'altre grup, totes les dades hi siguen superiors o iguals.

Dades no agrupades

Quan s'ordenen les dades, la posició que ocupa la mediana es determina dividint el nombre total de valors entre 2 ($\frac{n}{2}$) o el que és el mateix, calculant el 50% del total de dades ($0{,}5 \cdot n$). Cal tenir en compte, però, la paritat de n:

- Quan hi haja un nombre senar de valors, la mediana en serà just el valor central. Si per exemple es tenen les dades 3, 5, 5, 5, 6, 6, 7, la mediana és el valor 5, ja que hi ha tres valors més petits o iguals que 5, i tres valors més grans o iguals que 5.

  Si hi ha moltes dades, el càlcul no és immediat, cal construir la taula de freqüències i fixar-se en la columna de les freqüències absolutes acumulades N_i. La mediana serà el valor de variable que tinga la freqüència absoluta acumulada igual a $\frac{n}{2}$. És a dir:

  $$\text{si } N_{i-1} \leq \frac{n}{2} \leq N_i \rightarrow Me = x_i.$$

- Quan hi haja un nombre parell de valors, la mediana serà la mitjana aritmètica dels dos valors centrals de la variable. Si per exemple es tenen les dades $3, 5, 5, 5, 6, 6, 7, 8$ la mediana és el valor $\frac{5+6}{2} = 5{,}5$, ja que hi ha tres valors més petits o iguals a 5, i tres valors més grans o iguals que 6.

De la mateixa manera que en el cas anterior, si el conjunt d'observacions és nombrós, és necessari construir-ne la taula de freqüències i fixar-se en la columna de les N_i. Si en calcular $\frac{n}{2}$ aquest resulta ser un valor menor que una freqüència absoluta acumulada, la mediana es calcularà de la mateixa manera que en el cas anterior; és a dir, si $N_{i-1} \leq \frac{n}{2} \leq N_i \rightarrow Me = x_i$. Tanmateix, si coincideix $\frac{n}{2}$ amb algun N_i, per a obtenir-la es realitzarà el càlcul següent: $Me = \frac{x_i + x_{i+1}}{2}$. Els exemples següents en clarifiquen els càlculs.

Exemple 12

Es considera la variable X = nombre de feines diferents que han tingut els treballadors d'una empresa. La taula següent en recull la informació. Es demana que es calcule la mediana de la distribució de dades i que s'interprete l'estadístic.

x_i	n_i	N_i
1	3	3
2	4	7
5	9	16
7	**10**	**26**
10	7	33
13	2	35
	$n = 35$	

En primer lloc, cal calcular la posició que ocupa l'estadístic. En aquest cas, com que $\frac{n}{2} = \frac{35}{2} = 17{,}5$, la posició que ocupa és la 17,5. Així, tenint en compte el que hem explicat amb anterioritat i que si $N_{i-1} \leq \frac{n}{2} \leq N_i \rightarrow 16 \leq 17{,}5 \leq 26 \rightarrow Me = 7$.

Per altra part, el fet que la mediana siga 7 significa que a l'empresa hi ha almenys un 50% de treballadors que han ocupat 7 o menys feines diferents i almenys un 50% de treballadors que han ocupat 7 o més feines diferents.

Exemple 13

Es considera la variable X = nombre de feines diferents que han tingut els treballadors d'una empresa. La taula següent en recull la informació. Es demana que es calcule la mediana de la distribució de dades i que s'interprete l'estadístic.

x_i	n_i	N_i
1	3	3
2	4	7
5	**9**	**16**
7	10	26
10	6	32
	$n = 32$	

En primer lloc, cal calcular la posició que ocupa l'estadístic. En aquest cas, com que $\frac{n}{2} = \frac{32}{2} = 16$, la posició que ocupa és la 16. Així, tenint en compte el que s'ha explicat amb anterioritat i que $N_i = \frac{n}{2} \rightarrow 16 = \frac{32}{2} \rightarrow Me = \frac{x_i + x_{i+1}}{2} \rightarrow Me = \frac{5+7}{2} = 6$.

Dades agrupades

En distribucions agrupades, cal determinar l'interval $[L_{i-1},\ L_i)$ en el qual es troba la mediana. Aquest interval es determina seguint exactament els procediments esmentats en l'apartat anterior; es realitza el mateix que en el cas de dades no agrupades. La diferència radica que s'obtindrà un interval en lloc d'un valor.
Un colp es té l'interval $[L_{i-1},\ L_i)$, la mediana es calcula així:

$$Me = L_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} a_i, \text{ on,}$$

L_{i-1} És el límit inferior de la classe mediana.
N_{i-1} És la freqüència absoluta acumulada de la classe anterior a la classe mediana.
n_i És la freqüència de la classe mediana.
a_i És l'amplària de la classe mediana.

És evident que el que es pretén és calcular un representant de l'interval amb l'objecte de fixar la mediana en un valor. Una possibilitat hauria estat considerar la marca de classe; no obstant això, el criteri més seguit usualment no és aquest sinó el de la fórmula abans esmentada.

En aquesta fórmula, en primer lloc, es considera el supòsit que les dades estan uniformement distribuïdes dins de cada interval. Tenint en compte aquest fet, es pot observar que la fórmula és una relació de proporcionalitat entre les posicions que ocupen els valors de la variable i l'amplària dels intervals.

Exemple 14

La taula mostra la distribució dels sous d'una empresa en milers d'euros. Calcula'n la mediana.

$[L_{i-1}, L_i)$	n_i	N_i
[20 , 25)	100	100
[25 , 30)	150	250
[30 , 35)	**200**	**450**
[35 , 40)	180	630
[40 , 45]	41	671
	$n = 671$	

En primer lloc, cal calcular la posició en què es troba l'interval mediana.

$\frac{671}{2} = 335{,}5 \rightarrow$ L'interval és [30 , 35). Per tant:

$$Me = L_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} a_i \rightarrow Me = 30 + \frac{335{,}5 - 250}{200} \cdot 5 = 32{,}138$$ milers d'euros.

4.2.3. Mesures de localització centrals. Moda

És el valor de la variable que més vegades es repeteix, és a dir, el valor que té la freqüència absoluta més alta.

Poden existir distribucions amb més d'una moda: bimodals, trimodals, etc.

Dades no agrupades

En les distribucions sense agrupar, l'obtenció de la moda és immediata.

Exemple 15

La taula següent mostra el nombre de vehicles que tenen diferents empreses de la província de Castelló. Calcula'n la moda.

x_i	n_i
1	2
2	7
3	5
4	7
5	4

Moda {2, 4}, en aquest cas tenim una distribució bimodal.

Dades agrupades

En el supòsit que la distribució estiga donada en intervals, es poden produir dos casos: que tinguen la mateixa amplària, o que aquesta siga distinta. En ambdós casos l'objectiu es trobar un valor que represente la moda.

Intervals amb la mateixa amplària

És evident que una vegada determinada la freqüència més alta, a aquesta, no li correspon un valor sinó un interval. Llavors no tindrem un valor modal sinó un interval modal. Per a calcular el representant, de l'interval que faça el paper de moda hi ha diferents criteris. En el text es recull el següent. En primer lloc es calcula l'interval on es troba la moda, és a dir, l'interval modal $[L_{i-1}, L_i)$, el qual té la freqüència absoluta màxima (n_i). Posteriorment es calcula la moda de la manera següent:

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot a_i$$

On:

- L_{i-1}: extrem inferior de l'interval modal.
- a_i: amplitud d'aquest interval.
- n_{i-1}, n_{i+1}: freqüències dels intervals anterior i posterior, respectivament, de l'interval modal.

De la mateixa manera que la mediana, la fórmula té el supòsit que les dades estan uniformement repartides dins de cada interval. A més a més, seguint aquest criteri, es pot observar que la moda estarà més a prop d'aquell interval adjacent amb freqüència absoluta més alta.

Exemple 16

Trobar la moda de la següent distribució de dades referents a les puntuacions obtingudes pels aspirants a un lloc de treball en una prova de selecció. Calcula'n la moda.

$[L_{i-1}, L_i)$	n_i
[0, 25)	20
[25, 50)	140
[50, 75]	100

Es calcula la moda sabent que l'interval modal és [25, 50).

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot a_i \rightarrow Mo = 25 + \frac{100}{100 + 20} 25 = 45{,}83$$

Intervals amb amplària diferent

Quan els intervals siguen d'amplària distinta, l'interval modal serà el que tinga la densitat de freqüència, més alta ($d_i = \frac{n_i}{a_i}$), ja que es considerarà la qualitat de l'interval en funció de la freqüència i de l'amplària. Per a realitzar el càlcul:

$$Mo = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \cdot a_i.$$

El raonament que justifica la fórmula és evidentment el mateix que s'ha emprat amb la fórmula per als intervals amb la mateixa amplària.

Exemple 17

Troba la moda de la següent distribució de dades referents a les puntuacions obtingudes pels aspirants a un lloc de treball en una prova de selecció. Calcula'n la moda.

$[L_{i-1}, L_i)$	n_i	$d_i = n_i/a_i$
[0, 25)	20	0,8
[25, 50)	140	5,6
[50, 100)	180	3,6
[100, 150)	40	0,8
[150, 200]	20	0,4

Es calcula la moda sabent que l'interval modal és [25, 50):

$$Mo = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \cdot a_i \rightarrow Mo = 25 + \frac{3{,}6}{3{,}6 + 0{,}8} 25 = 45{,}53.$$

Nota: mitjana aritmètica, mediana i moda

Malgrat ser la mitjana aritmètica la mesura de tendència central més utilitzada, no sempre és la més adequada, a causa de la sensibilitat que presenta respecte de valors extrems o atípics. Aquest fet no afecta ni la mediana ni la moda. Tanmateix, en el càlcul de la mediana i la moda no influeixen totes les dades (com passa amb la mitjana aritmètica).

Cal dir que no existeixen criteris absoluts per a determinar quina mesura de tendència central és l'òptima en tots els casos. La conveniència d'una mesura o d'una altra dependrà de les dades i d'allò que representen.

4.2.4. Mesures de localització no centrals

Són mesures de localització semblants a la mediana. La seua funció és informar del valor de la variable que ocuparà la posició (en tant per cent) que ens interesse respecte de tot el conjunt d'observacions.

Podem dir que els quantils són unes mesures de posició que divideixen la distribució en un cert nombre de parts.

Les més importants són:

- Quartils. Divideixen la distribució en quatre parts iguals (tres divisions), Q_1, Q_2, Q_3, corresponents a 25%, 50%, 75%. Per exemple, el primer quartil té un 25% de les dades inferiors o iguals a ell, el segon quartil és la mediana, etc.
- Decils. Divideixen la distribució en deu parts iguals (nou divisions): D_1, ..., D_9, corresponents a 10%, ..., 90%.
- Percentils. Divideixen la distribució en cent parts (99 divisions): P_1, ..., P_{99}, corresponents a 1%, ..., 99%. Per exemple, el valor corresponent al 65è percentil, té un 65% de les dades inferiors o iguals a ell.

Hi ha un valor en el qual coincideixen els quartils, els decils i els percentils. És la mediana, ja que $P_{50} = Q_2 = D_5$. El càlcul dels quantils segueix el mateix procediment que el que s'ha utilitzat en la mediana, tant per a les dades agrupades com per a les dades sense agrupar. Així, en general, es calcula la posició en què es troba el quantil i després es calcula. Es distingeix entre distribucions agrupades i les que no ho estan:

Dades no agrupades

En primer lloc, es calcula la posició que ocupa el quantil que s'està calculant. Així, si Q_a representa el quantil que deixa per sota seu un a% de les dades:

si $N_{i-1} \leq \frac{a}{100} \cdot n \leq N_i \rightarrow Q_a = x_i$

en el supòsit que $\frac{a}{100} \cdot n = N_i \rightarrow Q = \frac{x_i + x_{i+1}}{2}$

Dades agrupades

En distribucions agrupades, cal determinar l'interval $[L_{i-1}, L_i)$ en el qual es troba el quantil. Aquest interval es determina seguint exactament els procediments esmentats en l'apartat anterior; es realitza el mateix que en el cas de dades no agrupades. La diferència radica que s'obtindrà un interval en lloc d'un valor. Un colp es té l'interval $[L_{i-1}, L_i)$, el quantil es calcula així:

$$Q_a = L_{i-1} + \frac{\%\cdot n - N_{i-1}}{n_i} a_i, \text{ on:}$$

L_{i-1} És el límit inferior de la classe on es troba el quantil.
N_{i-1} És la freqüència absoluta acumulada de la classe anterior a la classe on es troba el quantil.
n_i És la freqüència de la classe on es troba el quantil.
a_i És l'amplària de la classe on es troba el quantil.

Els exemples següents aclariran els conceptes i els procediments de càlcul.

Exemple 18

La taula següent mostra la distribució de les edats dels fills dels treballadors d'una empresa.

x_i	n_i	N_i
5	3	3
10	7	10
15	5	15
20	3	18
25	2	20
	$n = 20$	

Calcula la mediana (*Me*); el primer i el tercer quartil (Q_1, Q_3); el quart decil (D_4) i el norantè percentil (P_{90}).

*Mediana (*Me*)*

Lloc que ocupa la mediana al lloc $\frac{20}{2} = 10$è.

Com que és igual a un valor de la freqüència absoluta acumulada:

$$Me = \frac{10+15}{2} = 12{,}5.$$

Primer quartil (Q_1*)*

Lloc que ocupa en la distribució $\frac{1}{4}$ de $20 = \frac{20}{4} = 5$è.

Com que $N_{i-1} \leq 25\%$ de $\cdot\, n \leq N_i$, és a dir, $3 < 5 < 10$, llavors $Q_1 = x_i = 10$.

Tercer quartil (Q_3*)*

Lloc que ocupa en la distribució $\frac{3}{4}$ de $20 = \frac{60}{4} = 15$è, que coincideix amb un valor de la freqüència absoluta acumulada, per tant: $Q_3 = \frac{15+20}{2} = 17{,}5$.

Quart decil (D_4*)*

Lloc que ocupa en la distribució $\frac{4}{10}$ de $20 = \frac{80}{10} = 8$è.

Com que $N_{i-1} \leq 40\%$ de $\cdot\, n \leq N_i$, és a dir, $3 < 8 < 10$, per tant, $D_4 = 10$.

Norantè percentil (P_{90}*)*

Lloc que ocupa en la distribució 90% de 20 = 18è, que coincideix amb un valor de la freqüència absoluta acumulada, per tant:

$$P_{90} = \frac{20+25}{2} = 22{,}5.$$

Exemple 19

En la següent distribució de dades, troba el primer quartil, el quart decil i el norantè percentil:

$[L_{i-1}, L_i)$	n_i	N_i
[0, 100)	90	90
[100, 200)	140	230
[200, 300)	150	380
[300, 800)	120	500
	$n = 500$	

Primer quartil (Q_1*)*

Lloc que ocupa l'interval del primer quartil: $\frac{1}{4}$ de 500 = 125è. Per tant, Q_4 estarà situat a l'interval [100, 200). Aplicant-hi l'expressió directament:

$$Q_1 = 100 + \frac{125 - 90}{140} \cdot 100 = 125.$$

Quart decil (D_4*)*

Lloc que ocupa: $\frac{4}{410}$ de 500 = 200è. Per tant, D_4 estarà situat a l'interval [100, 200). Aplicant-hi l'expressió:

$$D_4 = 100 + \frac{200 - 90}{140} \cdot 100 = 178{,}57.$$

Norantè percentil (P_{90}*)*

Lloc que ocupa: 90% de 500 = 450è. Per tant, P_{90} estarà situat en l'interval [300, 800). Aplicant-hi l'expressió:

$$P_{90} = 300 + \frac{450 - 380}{120} \cdot 500 = 591{,}67.$$

4.3. Mesures de dispersió

És obvi que, amb l'objecte de descriure les dades, és de gran utilitat ubicar el centre del conjunt de dades. Però identificar una mesura de tendència central en rares ocasions és suficient. Una descripció més completa de les dades es pot obtenir si s'estudia com de disperses estan aquestes al voltant del punt central. Per exemple, la puntuació 4 és la mitjana aritmètica de les puntuacions obtingudes pel conjunt d'alumnes que van ser valorats amb {0, 0, 0, 0, 8, 8, 8, 8} i també del conjunt {4, 5, 6, 3, 2, 4, 5, 3}. És evident que la puntuació mitjana no representa amb la mateixa «fiabilitat» un grup d'alumnes que l'altre.

S'anomena *dispersió* o *variabilitat*, el grau de separació dels valors respecte de les mesures de centralització que s'estiguen tenint en compte. En conseqüència, les mesures de dispersió d'alguna manera informen de la representativitat de les mesures centrals respecte del conjunt de les dades.

En calcular una mesura de centralització com pot ser la mitjana aritmètica, resulta necessari acompanyar-la d'una altra mesura que indique el grau de dispersió de la distribució de la variable respecte d'aquesta mitjana aritmètica.

Aquesta quantitat o coeficient s'anomena *mesura de dispersió* i pot ser absoluta o relativa. Així doncs, les mesures de dispersió es poden classificar en:

Mesures de dispersió absolutes

- Recorregut
- Recorregut interquartílic
- Variància
- Desviació típica

Mesures de dispersió relatives

- Coeficient de variació de Pearson

4.3.1. Mesures de dispersió absolutes

Les mesures de dispersió que es tractaran tot seguit són el recorregut, el recorregut interquartílic, la variància i la desviació típica o estàndard.

Recorregut

Es defineix com la diferència entre els valors més gran i més petit de les variables d'una distribució de dades, és a dir:

$$Re = \text{màx.}(x_i) - \text{mín.}(x_i)$$

Recorregut interquartílic

Es defineix com la distància que hi ha entre el tercer i el primer quartil, és a dir:

$$Ri = Q_3 - Q_1.$$

Desviació mitjana respecte de la mediana

Es defineix com la mitjana aritmètica dels valors absoluts de les desviacions dels valors de la variable respecte de la mediana. Respon a l'expressió següent:

$$D_{|Me|} = \frac{\sum_{i=1}^{k} |x_i - Me| \cdot n_i}{n}.$$

Variància

Es defineix com la mitjana aritmètica dels quadrats de les desviacions dels valors de la variable respecte de la mitjana aritmètica de la distribució. Respon a l'expressió:

$$s^2 = \frac{(x_1 - \overline{X})^2 \cdot n_1 + (x_2 - \overline{X})^2 \cdot n_2 + \ldots + (x_k - \overline{X})^2 \cdot n_k}{n} = \frac{\sum_{i=1}^{k} (x_i - \overline{X})^2 \cdot n_i}{n}.$$

Com es pot observar en la definició, la variància és una mitjana del quadrat dels errors que es cometen en considerar la mitjana aritmètica com el representant de totes i cada una de les dades.

D'altra banda, una de les principals dificultats que presenta la variància és la unitat, ja que es mostra al quadrat (h^2, m^2, etc). La manera de solucionar-ho és calculant-ne l'arrel quadrada.

Desviació típica o desviació estàndard

Es defineix com l'arrel quadrada, amb signe positiu, de la variància. Respon a l'expressió següent:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \overline{X})^2 \cdot n_i}{n}}.$$

Nota

En les definicions anteriors s'han considerat dades no agrupades. Si ho foren, únicament caldria emprar les marques de classes com a representants dels intervals. És a dir, $c_i = x_i$.

Per altra part, es poden definir dos estadístics de dispersió més, anomenats *quasivariància* i *quasidesviació típica*, com:

$$s^2_{n-1} = \frac{n}{n-1} s^2 \qquad \text{i} \qquad s_{n-1} = \sqrt{\frac{n}{n-1}} \cdot s.$$

Aquests estadístics tenen molt d'interès en l'estadística inferencial, com es veurà en capítols posteriors.

Exemple 20

Es consideren les edats dels treballadors d'una gran empresa, les quals es mostren en la taula següent:

x_i	n_i
22	100
27	150
33	200
38	180
43	41
Total	671

Calcula'n el rang, el rang interquartílic, la desviació respecte de la mediana, la variància i la desviació típica.

Rang

El rang en aquest cas té un valor de $Re = 43 - 22 = 21$ anys. Aquest valor implica que no hi ha grans diferències d'edat entre els treballadors de l'empresa, si considerem que la diferència entre el treballador més jove i el més vell és de 21 anys.

Rang interquartílic

Per a calular el rang interquartílic cal, en primer lloc, calcular el tercer i el primer quartils. Fent els mateixos procediments que en l'apartat relatiu als quantils, s'obté que $Q_3 = 38$ i $Q_1 = 27$.

Per tant, el rang interquartílic és $Ri = 38 - 27$. Aquest valor implica que el 50% de les observacions centrals estan compreses en l'interval 27 – 38. Podem interpretar que si llevem el 25% dels treballadors més joves i el 25% dels més grans, el 50% restant té les edats compreses entre 27 i 38 anys.

Desviació respecte de la mediana, la variància i la desviació típica

Per a calcular aquest tres estadístics és necessari construir una taula semblant a la de les freqüències. Així:

x_i	n_i	$x_i n_i$	$(x_i - \overline{X})^2$	$(x_i - \overline{X})^2 \cdot n_i$	$\lvert x_i - Me \rvert \cdot n_i$
22	100	2.200	99,43	9.943,45	1.100
27	150	4.050	24,72	3.707,65	900
33	200	6.600	1,06	211,49	0
38	180	6.840	36,34	6.541,31	900
43	41	1.763	121,62	4.986,57	410
Total	671	21.453		25.390,46	3.310

Emprant els procediments dels epígrafs anteriors s'obté que la mediana és 33 i la mitjana aritmètica és 31,97.

Aleshores:

La desviació mitjana respecte de la mediana: $D_{|Me|} = \dfrac{\sum_{i=1}^{k} |x_i - Me| n_i}{n} = \dfrac{3310}{671} = 4{,}93$ anys.

La variància és $s^2 = \dfrac{\sum_{i=1}^{k} (x_i - \overline{X})^2 \cdot n_i}{n} = \dfrac{25390{,}46}{671} = 37{,}84$ anys^2.

La desviació típica és $s = \sqrt{s^2} = \sqrt{37{,}84} = 6{,}15$ anys.

Nota

El càlcul de la variància també es pot realitzar d'una altra manera. Únicament cal tenir en compte que $(x_i - \overline{X})^2 = x_i^2 - 2x_i\overline{X} + \overline{X}^2$, i realitzar alguns càlculs algebraics per a arribar-ne a la segona expressió:

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{X})^2 \cdot n_i}{n} = \frac{\sum_{i=1}^{k} x_i^2 \cdot n_i}{n} - \overline{X}^2.$$

Aquesta darrera expressió és la més utilitzada per a realitzar el càlcul. No obstant això, cal construir una taula de freqüències diferent per a utilitzar aquesta segona expressió.

Encara que el càlcul manual de la variància és senzill, de la mateixa manera que per a la mitjana aritmètica, cal dir que quasi totes les calculadores científiques admeten un procediment molt senzill en el paquet estadístic, que permet calcular-la amb molta facilitat. Evidentment, també s'hi poden emprar els fulls de càlcul i els programes informàtics estadístics.

Propietats de la variància

La variància compleix les propietats següents:

Propietat 1
La variància és sempre un valor no negatiu o 0. Únicament pot ser 0 si totes les dades són iguals. En aquest cas és evident que $\overline{X} = x_i$ *per a tots els possibles valors de l'índex.*

La demostració és evident a partir de la definició de *variància*.

Propietat 2
Si a tots els valors de la variable se suma una constant, la variància no es modifica.

La demostració és relativament senzilla des del punt de vista algebraic.

Si es construeix la variable *X'* sumant a cada valor de la variable *X* una constant *k*, llavors la transformació de cada dada és $x_i' = x_i + k$.

D'altra banda, aplicant les propietats de la mitjana aritmètica se sap que $\overline{X'} = \overline{X} + k$.

Així doncs:

$$s'^2 = \frac{\sum_{i=1}^{k}(x'_i - \overline{X'})^2 \cdot n_i}{n} = \frac{\sum_{i=1}^{k}\left((x_i + k) - \left(\overline{X} + k\right)\right)^2 \cdot n_i}{n} = \frac{\sum_{i=1}^{k}(x_i - \overline{X})^2 \cdot n_i}{n} = s^2$$

És a dir:

El fet que una variable es genere a partir d'una altra sumant-li una constant ($X' = X + K$) significa el que es mostra tot seguit:

Valors de la variable inicial	*Valors de la variable modificada*
x_1	$x'_1 = x_1 + k$
x_2	$x'_2 = x_2 + k$
....	.....
x_k	$x'_k = x_k + k$

La propietat afirma que, si això és així, llavors $s'^2 = s^2$.

Exemple 21

Els sous de l'any 2006 de 20 treballadors d'una empresa es mostren en la taula següent:

***X* = sous (€)**	**Treballadors**
17.500	12
22.500	6
27.500	1
32.500	1

La mitjana aritmètica dels sous és 20.250 € i la variància, 4.023,69 €^{2}.

Si l'any 2007 l'empresa decideix apujar el sou de cada treballador 1.000 €, llavors s'obté una nova variable *X'* (*X'* = sou dels treballadors per a l'any 2007) a partir de l'anterior. L'equació que relaciona totes dues variables és:

$$X' = X + 1.000.$$

Sous any 2007 $\rightarrow X' = X + 1.000$

Aplicant-hi la propietat s'obté que la variància de *X'*, és la mateixa que la de *X*, és a dir, 4.023,69 €^{2}.

Propietat 3
Si tots els valors de la variable es multipliquen per una constant, la variància queda multiplicada pel quadrat de la dita constant:

La demostració és relativament senzilla des del punt de vista algebraic.

Si es construeix la variable *X'* multiplicant cada valor de la variable *X* per una constant *k*, llavors la transformació de cada dada és $x_i' = k \cdot x_i$.

D'altra banda, aplicant-hi les propietats de la mitjana aritmètica se sap que $\overline{X}' = k \cdot \overline{X}$.

Així doncs:

$$s'^2 = \frac{\sum_{i=1}^{k}(x'_i - \overline{X'})^2 \cdot n_i}{n} = \frac{\sum_{i=1}^{k}\left((k \cdot x_i) - \left(k \cdot \overline{X}\right)\right)^2 \cdot n_i}{n} = \frac{\sum_{i=1}^{k}\left(k(x_i - \overline{X})\right)^2 \cdot n_i}{n}$$

$$= k^2 \frac{\sum_{i=1}^{k}(x_i - \overline{X})^2 \cdot n_i}{n} = k^2 \cdot s^2$$

És a dir:

El fet que una variable es genere a partir d'una altra multiplicant-la per una constant ($X' = a \cdot X$) significa el que es mostra tot seguit:

Valors de la variable inicial	*Valors de la variable modificada*
x_1	$x'_1 = a \cdot x_1$
x_2	$x'_2 = a \cdot x_2$
....	.....
x_k	$x'_k = a \cdot x_k$

La propietat afirma que, si això és així, llavors $s'^2 = a^2 \cdot s^2$.

Exemple 22

Es consideren les mateixes dades que a l'exemple anterior, és a dir, *X* = sous de 20 treballadors l'any 2006.

Se suposa que l'empresari decideix apujar el sou de cada treballador per a l'any 2007 un 2%. D'aquesta manera sorgeix una altra variable *X'* (*X'* = sou dels treballadors l'any 2007) en funció de la variable *X*. L'equació que relaciona totes dues variables és $X' = 1{,}02 \cdot X$ (apujar una quantitat un 2% és equivalent a multiplicar-la per 1,02).

Per a calcular la variància de la variable *X'* (sou mitjà l'any 2007) no cal construir la taula de freqüències de la variable *X'*, tan sols aplicar-hi la propietat.

Així:

$$s'^2 = a^2 \cdot s^2 \rightarrow s'^2 = 1{,}02^2 \cdot s^2 = 1{,}02 \cdot 4023{,}69 = 4.186{,}24\ €^2.$$

Propietat 4
Si una variable X'*és transformació lineal d'una altra variable* X *(*X' = a · X + b*; a i b nombres reals), la variància de* X' *s'obté a partir de la de* X *de la manera* $s'^2 = a^2 \cdot s^2$.

La demostració és relativament senzilla des del punt de vista algebraic.

Si es construeix la variable *X'* multiplicant cada valor de la variable *X* per una constant *k*, llavors la transformació de cada dada és $x_i' = a \cdot x_i + b$.

D'altra banda, aplicant-hi les propietats de la mitjana aritmètica se sap que $\overline{X}' = a \cdot \overline{X} + b$.

Així doncs:

$$s'^2 = \frac{\sum_{i=1}^{k} (x'_i - \overline{X}')^2 \cdot n_i}{n} = \frac{\sum_{i=1}^{k} \left((a \cdot x_i + b) - (a \cdot \overline{X} + b)\right)^2 \cdot n_i}{n} = \frac{\sum_{i=1}^{k} \left(a(x_i - \overline{X})\right)^2 \cdot n_i}{n}$$

$$= a^2 \frac{\sum_{i=1}^{k} (x_i - \overline{X})^2 \cdot n_i}{n} = a^2 \cdot s^2$$

És a dir:

El fet que una variable es genere a partir d'una altra multiplicant-la per un nombre i sumant-li una constant ($X' = a \cdot X + b$) significa el que es mostra tot seguit:

Valors de la variable inicial	*Valors de la variable modificada*
x_1	$x'_1 = a \cdot x_1 + b$
x_2	$x'_2 = a \cdot x_2 + b$
....	.....
x_k	$x'_k = a \cdot x_k + b$

La propietat afirma que, si això és així, llavors $s'^2 = a^2 \cdot s^2$. És a dir, la suma de la constant no hi influeix.

Exemple 23

Es consideren les mateixes dades que a l'exemple de la segona propietat, és a dir, X = sous de 20 treballadors l'any 2006.

Se suposa ara que l'empresari decideix apujar el sou de cada treballador per a l'any 2007 un 2% i, a més, 1.000 € en concepte de prima. D'aquesta manera sorgeix una altra variable $X' = 1{,}02X + 1.000$.

Per a calcular la variància de la variable X' (sou mitjà l'any 2007) no cal construir la taula de freqüències, tan sols aplicar-hi la propietat.

Així:

$$s'^2 = a^2 \cdot s^2 \qquad s'^2 = 1{,}02^2 \cdot s^2 = 1{,}02 \cdot 4023{,}69 = 4.186{,}24\ €^2$$

Propietat 5
Si d'un conjunt de valors es poden obtenir dos o més subconjunts disjunts que formen una partició d'aquest conjunt, la variància de tot el conjunt de dades està relacionada amb les variàncies dels subconjunts.

Així, si es tenen n dades distribuïdes en k subconjunts disjunts de grandàries

$N_1, N_2, ..., N_k$ i $\sum N_i = n \rightarrow$

$$\rightarrow S_2 = \frac{(S^2_1 + \bar{x}^2_1)N_1 + (S^2_2 + \bar{x}^2_2)N_2 + ... + (S^2_k + \bar{x}^2_k)N_k}{n} - \left(\frac{\bar{x}_1 N_1 + \bar{x}_2 N_2 + ... + \bar{x}_k N_k}{n}\right)^2$$

On: $\bar{x}_i \rightarrow$ mijana aritmètica del subconjunt i
$S^2_1 \rightarrow$ variància del subconjunt i

És a dir:

Aquesta propietat permet calcular la variància total si es coneixen les variàncies parcials.

Exemple 24

S'està estudiant el nombre de persones que comparteixen pis d'una mostra dels estudiants de l'UJI. Per a fer-ho, s'escullen 200 alumnes de l'Escola Superior de Tecnologia i Ciències Experimentals, 150 alumnes de la Facultat de Ciències Jurídiques i Econòmiques, i 250 de la Facultat de Ciències Humanes i Socials. Sobre aquestes dades s'han calculat els estadístics mitjana aritmètica i variància. Es desitja conèixer la mitjana aritmètica i la variància respecte dels 600 alumnes enquestats.

Les dades recollides es mostren en la taula següent:

ESTC	$\overline{x_{ESTC}} = 4{,}25 \quad S^2_{ESTC} = 1{,}025$
FCJE	$\overline{x_{FCJE}} = 3{,}95 \quad S^2_{ESTC} = 0{,}925$
FCHS	$\overline{x_{FCHS}} = 5{,}02 \quad S^2_{FCHS} = 0{,}75$

En aquest cas els conjunts disjunts estan formats pels enquestats en cada facultat i escola. De cada subconjunt es coneixen la variància i la mitjana aritmètica; per tant, és possible aplicar-hi la propietat.

$$S^2 = \frac{(1{,}025 + 18{,}0625)200 + (0{,}925 + 15{,}6025)150 + (0{,}75 + 25{,}004)250}{600} - \left(\frac{850 + 592{,}5 + 1255}{600}\right)^2$$

$$= 0{,}8215$$

Per a calcular la mitjana aritmètica únicament cal aplicar-hi la propietat adient de la mitjana aritmètica. Aplicant-la s'obté que:

$$\bar{x} = \frac{200 \cdot 4{,}25 + 150 \cdot 3{,}95 + 250 \cdot 5{,}02}{600} = 4{,}4958.$$

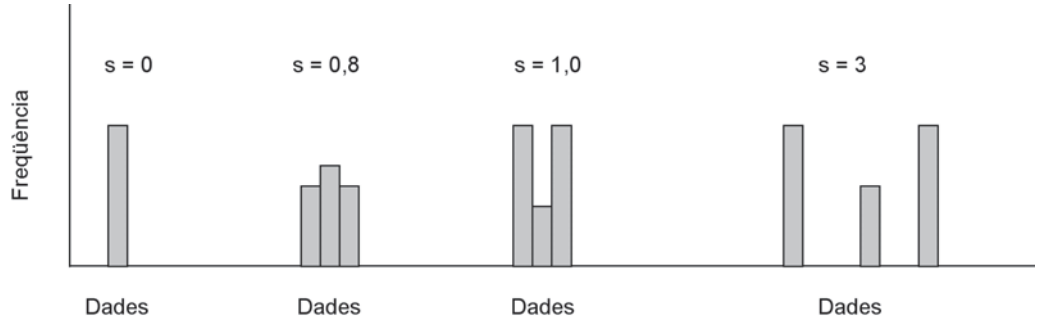
Propietats de la desviació típica

La desviació també té una sèrie de propietats que es dedueixen directament de les de la variància (ja que la desviació típica és l'arrel quadrada de la variància):

- La desviació típica és sempre un valor no negatiu. S serà sempre superior o igual a 0 per definició. $S = 0$ quan $X = x_i$ (per a tot i).
- Si a tots els valors de la variable, se'ls suma una mateixa constant, la desviació típica no varia. És a dir, si $X' = X + k$, aleshores $s' = s$.
- Si tots els valors de la variable es multipliquen per una mateixa constant, la desviació típica queda multiplicada pel valor absolut de la dita constant. És a dir, si $X' = a \cdot X$, aleshores $s' = |a| \cdot s$.
- Si una variable X' és transformació lineal d'una altra variable X ($X' = a \cdot X + b$; a i b nombres reals), la desviació típica de X' s'obté a partir de la de X aplicant-hi que: $s' = |a| \cdot s$.

Interpretació de la desviació estàndard

Com ja s'ha comentat, la desviació típica és útil per a descriure un conjunt de dades mesurant el grau de dispersió de les dades al voltant de la mitjana aritmètica. Si les observacions estan molt pròximes entre si, la mitjana aritmètica serà representativa i la desviació estàndard serà petita. Si, al contrari, les dades estan més separades entre si, més disperses, la desviació típica serà més gran. És a dir, com més disperses estan les observacions, més gran és el valor de la desviació típica (observeu el gràfic següent).



Atès que la variació típica és una eina útil per a mesurar la variació, es consideraran dues formes (no en són les úniques) per a aconseguir una comprensió intuïtiva de la desviació estàndard, així com de la seua utilitat.

Teorema de Txebixev

El teorema de Txebixev fou formulat pel matemàtic rus P. L. Txebixev (1821-1894). Estableix que per a qualsevol conjunt de dades, almenys el $\left(1-\frac{1}{K^2}\right)$% de les observacions són dins d'un interval que té per centre la mitjana aritmètica i per radi, k desviacions de la mitjana aritmètica, on k és qualsevol nombre superior a 1.

Per exemple, si es pren $k = 3$, a l'interval $[\overline{X} - 3s,\ \overline{X} + 3s]$ pertanyen el $\left(1-\frac{1}{3^2}\right)$ = 88,89% de les dades.

Exemple 25

Se suposa un conjunt d'observacions referents a les hores reals que treballen a la setmana persones d'empreses d'una província. De les dades, únicament es coneix que la mitjana aritmètica és de 47 hores i la desviació típica, de 4 hores.

Aplicant el teorema de Txebixev, es pot saber que el 75% de les observacions estan compreses entre 39 i 55 hores. [Es pren $k = 2 \rightarrow \left(1 - \frac{1}{K^2}\right) = 0{,}75$. Llavors $[\overline{X} - 2s,\ \overline{X} + 2s] = [39{,}55]$].

Cal recordar que aquesta regla únicament té en compte la mitjana aritmètica i la desviació típica; no té en compte ni el nombre de dades, ni si les freqüències de les dades segueixen una distribució determinada, etc. Els resultats, doncs, són molt aproximats i, en moltes ocasions, poc efectius.

La distribució normal i la regla empírica

En estadística se sol dir que la representació gràfica de les dades aporta molta informació al voltant de com es distribueixen les observacions, és a dir, de si aquestes estan molt disperses o, al contrari, estan agrupades al voltant de la mitjana aritmètica. En particular, de l'observació dels histogrames, es poden classificar els conjunts de dades segons uns models de distribucions de dades teòriques. És a dir, si l'histograma d'un conjunt d'observacions té la forma semblant a un d'aquests models teòrics, llavors aquest fet permet suposar que les dades segueixen el mateix patró teòric.

Però com s'obtenen aquests patrons teòrics? N'hi ha diferents maneres, però una molt intuïtiva consisteix a prendre mostres de grandària considerable, representar l'histograma i buscar una funció matemàtica que s'ajuste tant com siga possible a les freqüències relatives de cadascun dels intervals. Així, per exemple, la figura 4 representa aquest estudi per a una mostra de 3.000 observacions de la variable estadística X = metres quadrats dels habitatges d'una determinada ciutat. En el primer gràfic es mostra l'histograma amb 10 intervals. En el segon gràfic l'histograma té molts més intervals, ja que del que es tracta és de trobar una funció que s'ajuste al màxim a les freqüències relatives dels intervals, i amb més intervals a l'histograma, l'ajustament és més precís. En el tercer histograma apareixen també representades diferents funcions matemàtiques, que són les candidates a ser el model teòric que segueixen les freqüències relatives de les dades. Per a acabar, el quart gràfic representa el model teòric que més s'hi ajusta. La interpretació del model teòric permet fer afirmacions com ara que un percentatge molt alt de les observacions es troben al voltant dels 200 m^2; és més, la gran majoria es troben compreses entre 196 m^2 i 204 m^2. Cal notar que per a fer afirmacions més contundents com ara que la mitjana aritmètica de la superfície dels habitatges de tota la població d'estudi (cal recordar que estavem parlant d'una mostra de 3.000 dades) és de 200 m^2 o que la variable en la població que ha originat la mostra també segueix aquest patró, caldria fer-ne un estudi estadístic més profund.

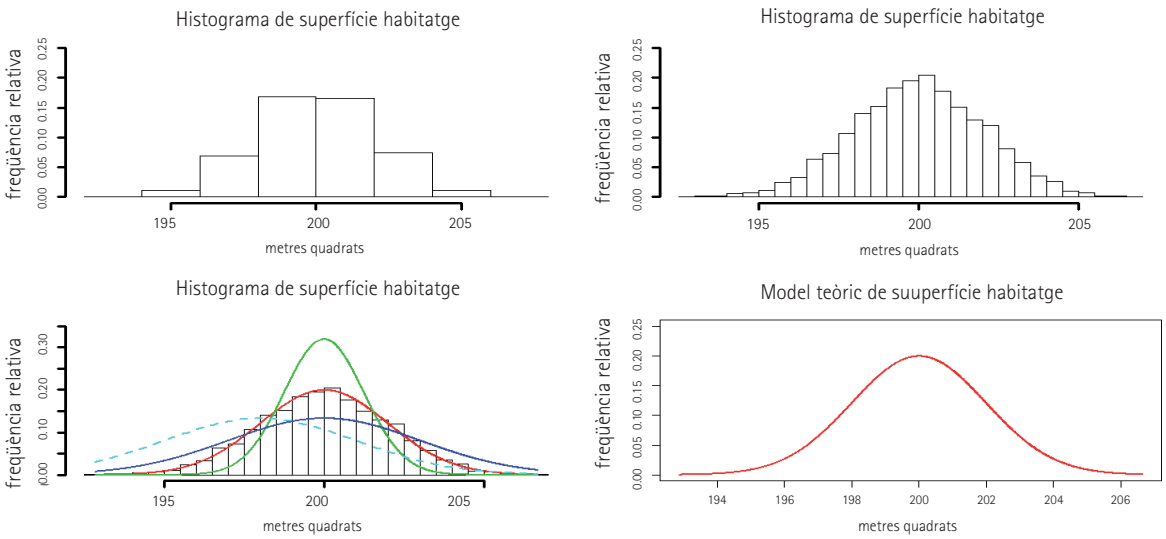


Figura 4

Per altra banda, de tots els models teòrics, el més important és l'anomenat *model normal* o *distribució normal*. La funció matemàtica que determina aquest model teòric és:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Té forma de campana invertida (l'elecció de l'exemple anterior no ha estat aleatòria) i, com es pot observar, està en funció de dos paràmetres, la mitjana (μ) teòrica i la desviació típica (σ) teòrica. Per a cada model teòric concret, els valors dels paràmetres serien els que s'obtindrien en una població teòrica adequada a aquests valors. A més a més, la mitjana aritmètica i la desviació típica determinen completament el model teòric, i el diferencien d'altres models normals amb diversos valors dels paràmetres. A la figura 5 es poden observar diferents conjunts de dades que segueixen models normals amb valors dels paràmetres distints. Així, es pot observar que tots els histogrames tenen una forma semblant i, en conseqüència, els patrons teòrics també han de ser-ho. En aquest cas tots quatre tenen forma de campana invertida (campana de Gauss). Les diferències radiquen que quan la desviació típica augmenta, la campana s'aplana. Tanmateix, en totes tres, les dades estan distribuïdes de manera semblant al voltant de la mitjana aritmètica: la freqüència relativa de les dades que són properes a la mitjana aritmètica és més gran que la de les que hi són més lluny.

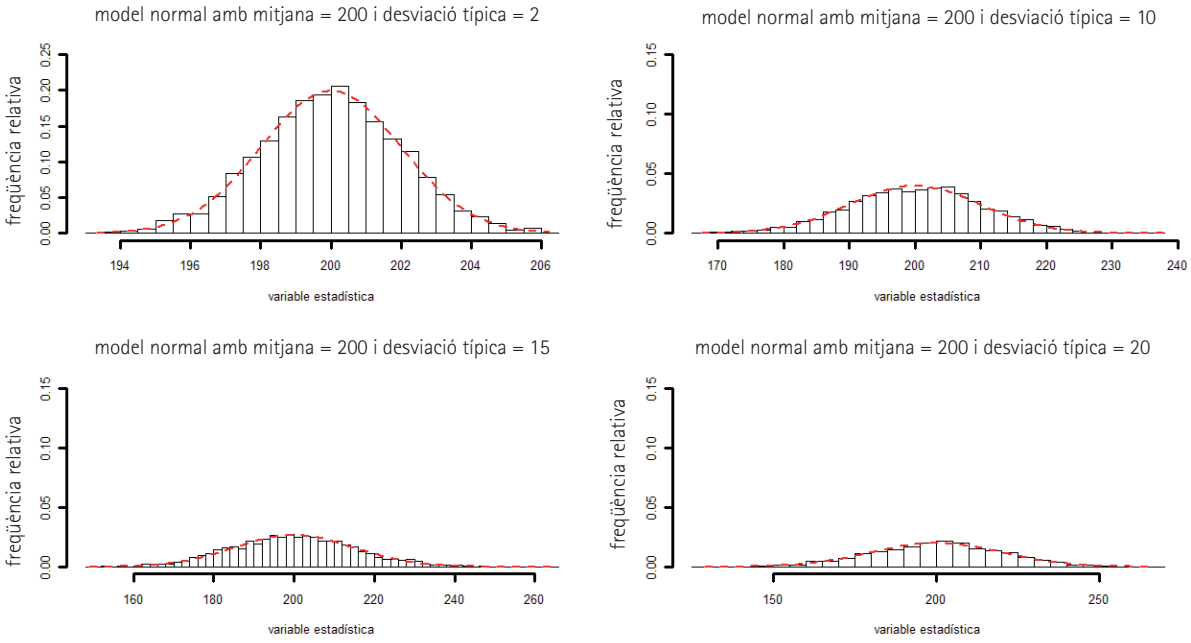


Figura 5

Malgrat que no és ara el moment d'aturar-se per a realitzar l'estudi detallat del model normal –estudi que es realitzarà en unitats posteriors–, sí que és convenient donar una sèrie de propietats sobre aquest model amb l'objecte de poder aplicar el que Allen L. Webster (2000) anomena *regla empírica*, i observar la utilitat de la desviació estàndard per a realitzar aproximacions semblants a les del teorema de Txebixev.

Cal recordar que, per a poder aplicar aquest teorema al conjunt d'observacions, no se'ls exigia res, i en conseqüència, els resultats no eren massa efectius. Tanmateix, si el conjunt d'observacions segueixen un patró normal, aleshores els resultats que se'n poden traure són molt més interessants.

Així, tenint en compte que tot model teòric normal produeix una corba simètrica en forma de campana; que aquest model queda determinat, com ja s'ha esmentat, per dos paràmetres (la mitjana aritmètica (μ) i la desviació típica (σ)); i que per a qualsevol model normal es compleix que:

- el 68,3% de les observacions pertany a l'interval $[\mu - \sigma, \mu + \sigma]$
- el 95,5% de les observacions pertany a l'interval $[\mu - 2 \cdot \sigma, \mu + 2 \cdot \sigma]$
- el 99,7% de les observacions pertany a l'interval $[\mu - 3 \cdot \sigma, \mu + 3 \cdot \sigma]$;

la regla empírica afirma que si un conjunt d'observacions segueix un patró normal, aleshores el 68,3% de totes les observacions estaran a una desviació típica de la mitjana. Si en lloc de considerar una desviació típica per damunt i per davall, es consideren dues o tres desviacions, el percentatge de dades dins dels intervals augmenta, i són, respectivament, 95,5% i 99,7%. A més a més, cal notar que aquests percentatges no depenen dels valors, ni de la mitjana, ni de la desviació típica, sinó únicament de si les dades segueixen un model normal o no.

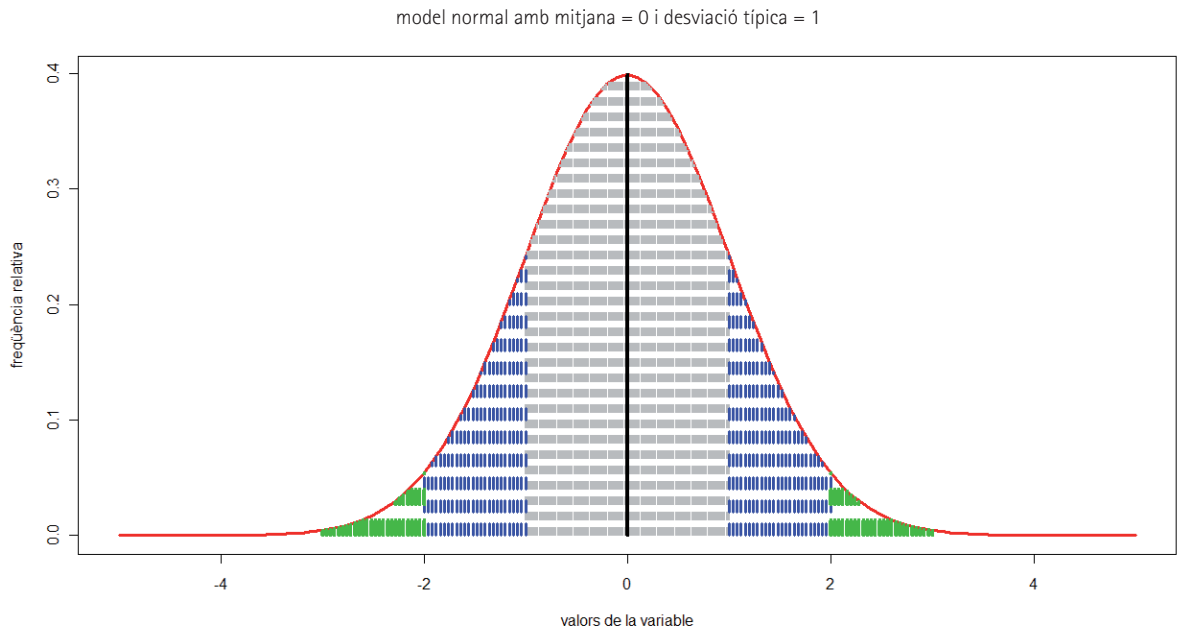


Figura 6

La figura 6 mostra una variable que segueix una distribució normal, de mitjana aritmètica 0 i desviació típica 1. A l'eix vertical s'hi representa la freqüència relativa.

Per tant, es pot comprovar gràficament que:

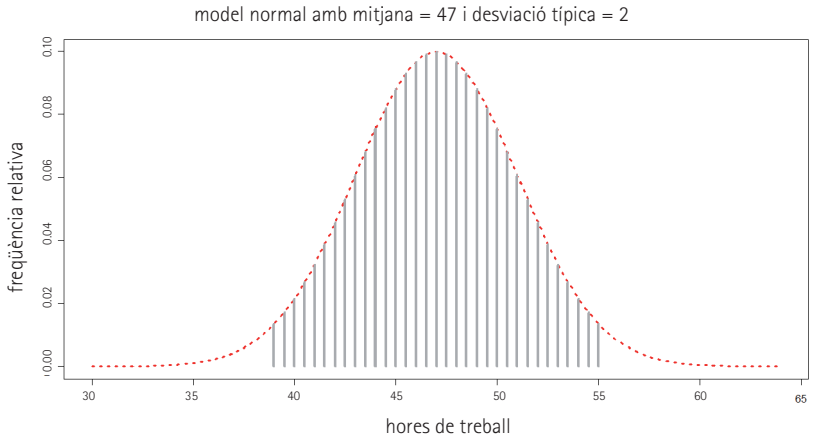
- el 68,3% de les observacions pertany a l'interval [–1, 1] (gris)
- el 95,5% de les observacions pertany a l'interval [–2, 2] (gris + blau)
- el 99,7% de les observacions pertany a l'interval [–3, 3] (gris + blau + verd)

També cal dir que, a mesura que la desviació típica va augmentant (augment de la dispersió), la campana es va xafant. No obstant això, totes les propietats anteriors se segueixen complint.

Per altra part, l'objecte d'aquest breu estudi no és altre que el d'observar la utilitat de la desviació típica. Així doncs, si l'histograma de freqüències d'un conjunt de dades té una forma semblant a la d'una campana com la de la figura, aleshores es pot suposar que les dades segueixen una distribució de freqüències normal i s'hi pot aplicar la regla empírica.

Exemple 26

Se suposa un conjunt d'observacions referents a les hores reals que treballen a la setmana persones d'empreses d'una província. De les dades, únicament es coneix que la mitjana aritmètica és de 47 hores i la desviació típica, de 4 hores. A més a més, s'ha observat que l'histograma de freqüències absolutes és molt semblant a la campana de Gauss.



Així, aplicant l'anterior, es pot saber que el 95,5% de les dades estan compreses entre 39 i 55 hores.

Com que es pot suposar que les dades segueixen un model normal, llavors el 95,5% de les dades pertany a l'interval $[\overline{X} - 2s,\ \overline{X} + 2s] = [39, 55]$.

4.3.2. Mesures de dispersió relatives

Les mesures de dispersió absolutes són uns indicadors que presenten dificultats a l'hora de comparar la representativitat de les mesures de tendència central entre dues distribucions de dades diferents. Per això, a vegades es recorre a mesures de dispersió relatives. El coeficient de variació de Pearson n'és una de les més significatives i determina el grau de representativitat de la mitjana aritmètica relativa al conjunt de dades que representa. Es defineix com el quocient entre la desviació típica i la mitjana aritmètica de la distribució de dades: $V_X = \frac{s}{|\overline{X}|}$.

És necessari tenir en compte que en efectuar el quocient s'eliminen les unitats. Consegüentment, V_X és adimensional. A més a més, atès que la desviació típica és una espècie de mitjana de l'error, per considerar la mitjana com el representant de tots i cadascun dels valors de la variable, el quocient entre la desviació i la mitjana aritmètica es pot considerar com la proporció d'error que té la mitjana.

Així doncs, quan $Vx < Vy$ significa que $\overline{X}$ és més representativa de les dades que $\overline{Y}$, o dit d'una altra manera, la distribució X és més homogènia que Y. Per convenció es considera que la dispersió és òptima i, per tant, més homogènia si Vx és igual o inferior a 0,3.

Exemple 27

Per a l'empresa A el nombre mitjà de comandes mensuals d'un determinat producte és de 250 i la desviació típica, de 12. Per a l'empresa B el nombre mitjà de comandes és de 375 amb una desviació típica de 15. Quina de les dues empreses fa unes comandes mensuals més homogènies?

Per a contestar a la pregunta cal calcular l'índex de variació d'ambdues companyies.

$$V_{X_A} = \frac{s_A}{|\overline{X_A}|} = \frac{12}{250} = 0{,}048 \; V_{X_B} = \frac{s_B}{|\overline{X_B}|} = \frac{15}{375} = 0{,}040$$

Com que l'índex de variació de la companyia A és superior al de la companyia B, les seues comandes mensuals són més disperses.

4.4. Mesures de forma i diagrama de caixa

Fins ara, per a analitzar les dades s'han emprat, sobretot, les mesures de centralització i de dispersió, les quals aporten molta informació al voltant de la distribució de dades. No obstant això, encara hi ha algunes qüestions que no es poden obtenir a partir d'aquests estadístics. Per exemple, és possible que distribucions amb els mateixos valors dels estadístics mitjana aritmètica i desviació típica tinguen representacions gràfiques diferents (figura 7) és a dir, distribucions en què les freqüències dels valors que pren la variable poden ser molt diferents i, per contra, també és possible tenir la mateixa dispersió i la mateixa mitjana aritmètica. És per això que l'observació de les representacions gràfiques de les dades completa el coneixement que s'hi pot obtenir.

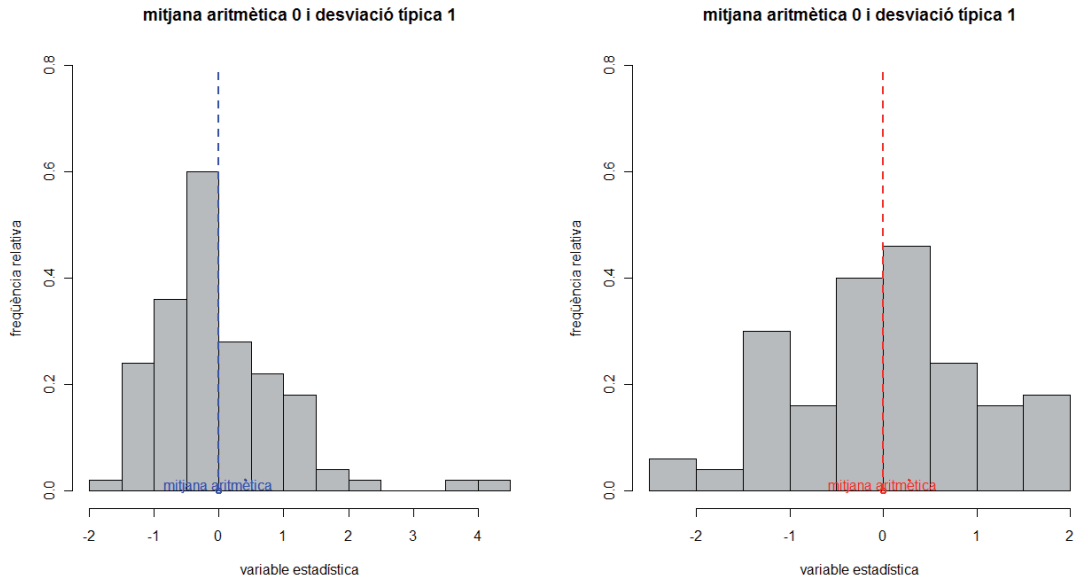


Figura 7

En aquest epígraf el que estudiarem són dos estadístics que permeten saber la forma dels gràfics de les dades sense fer-ne la representació. Aquestes mesures s'anomenen *mesures de forma* i es classifiquen en dos grans grups: mesures d'asimetria i mesures de curtosi. Per a finalitzar aquest apartat s'introduirà el diagrama de caixa, el qual resumeix molta de la informació que mostren els diferents estadístics esmentats fins ara.

4.4.1. Mesures d'asimetria

Les mesures d'asimetria permeten determinar, sense que siga necessari fer les representacions gràfiques, el grau de simetria que presenten les dades respecte d'un valor central de la variable estadística, normalment la mitjana aritmètica. Per tant, aquesta mesura ha de reflectir dos aspectes: la distància de cada observació respecte de la mitjana aritmètica, és a dir, la diferència entre cada valor i la mitjana aritmètica: $(x_i - \overline{x})$, i la freqüència de cadascuna d'aquestes distàncies (la qual coincidirà, evidentment, amb la freqüència de cada observació). D'aquesta manera, intuïtivament, si hi predominen les distàncies negatives sobre les positives (per ser més freqüents o ser distàncies molt grans), llavors la distribució és asimètrica per l'esquerra. Si, al contrari, es dóna la situació oposada, llavors la distribució és asimètrica per la dreta. Per a finalitzar, si les distàncies negatives i les positives es compensen, llavors la distribució és simètrica (figura 8).

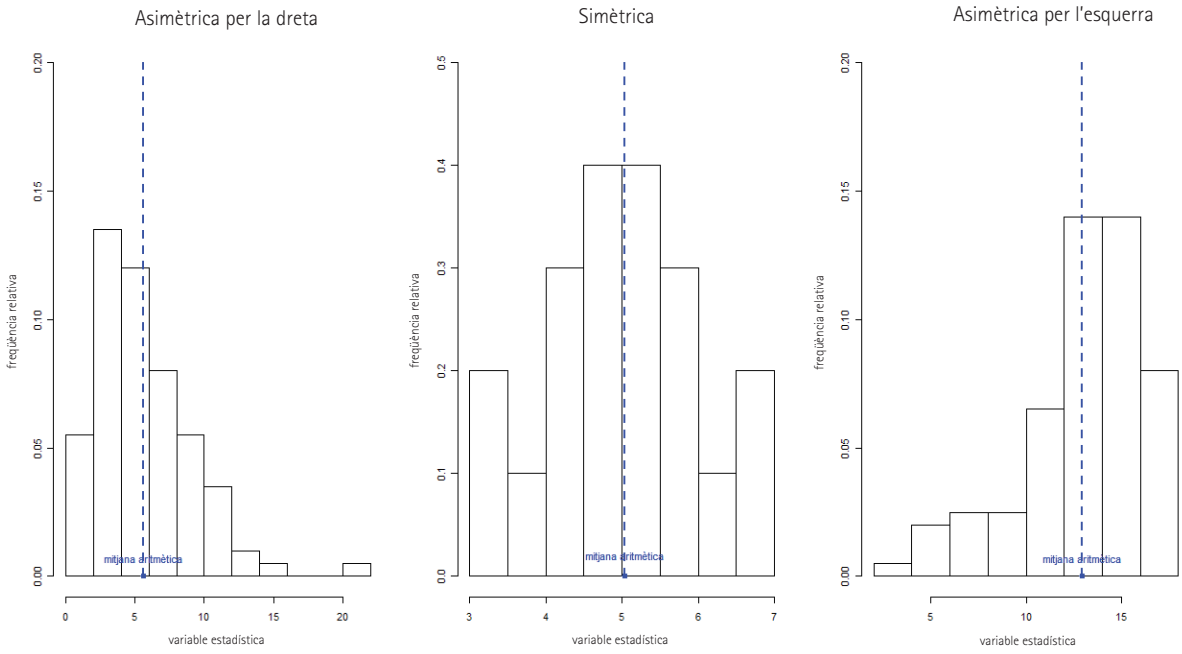


Figura 8

Ara doncs, el que cal és trobar l'estadístic que determine l'asimetria de la distribució de dades. Com que l'asimetria està directament relacionada amb les desviacions respecte de la mitjana aritmètica, un primer apropament pot ser la mitjana de les desviacions, és a dir $\frac{\sum_{i=1}^{k}(x_i - \overline{X})n_i}{n}$. No obstant això, ja és conegut que aquesta suma és 0 (propietats de la mitjana aritmètica).

D'altra banda, com que ens interessa conèixer el signe de les desviacions, tampoc podem emprar el quadrat de les desviacions. Així doncs, sembla coherent prendre una potència de grau 3 de les desviacions i calcular-ne la mitjana. Així, si anomenem

$$m = \frac{\sum_{i=1}^{k}(x_i - \overline{X})^3 n_i}{n},$$

llavors es compleix que:

si $m_1 = 0$	la distribució és simètrica,
si $m_1 > 0$	la distribució és asimètrica positiva,
si $m_1 < 0$	la distribució és asimètrica negativa.

Cal tenir en compte que el fet d'elevar al cub provoca que distàncies inferiors a 1 es facen més menudes ($0{,}1^3 = 0{,}001$) i que distàncies superiors a 1 augmenten molt més ($4^3 = 64$). És per això que si hi ha dades distanciades, aquestes tenen molt de pes dins de l'estadístic. A més a més, pareix lògic que si la suma de les desviacions positives és més gran que la de les desviacions negatives, la distribució de dades siga asimètrica per la dreta o positiva. Si es dóna el cas oposat, les desviacions negatives més grans que les positives, la distribució serà asimètrica per l'esquerra o negativa.

Per altra part, aquest indicador té per dimensió el cub de la dimensió de la variable estudiada, la qual cosa sol ocasionar problemes a l'hora de fer canvis d'escala. Per a aconseguir un indicador sense dimensió cal dividir l'expressió anterior per una quantitat que estiga donada en les mateixes unitats. Aquesta quantitat és el cub de la desviació típica. D'aquesta manera s'obté el coeficient d'asimetria de Fisher:

$$g_1 = \frac{m}{s^3} = \frac{\dfrac{\sum_{i=1}^{k}(x_i - \overline{X})^3 n_i}{n}}{\left(\sqrt{\dfrac{\sum_{i=1}^{k}(x_i - \overline{X})^2 n_i}{n}}\right)^3}.$$

Cal notar que com que la desviació típica és positiva, el signe del coeficient de Fisher serà el mateix que el de m. I, per tant:

si $g_1 = 0$	la distribució és simètrica,
si $g_1 > 0$	la distribució és asimètrica positiva,
si $g_1 < 0$	la distribució és asimètrica negativa,

Així doncs, quan $g_1 < 0$, es diu que la distribució presenta asimetria per l'esquerra (o negativa) i llavors, de les dues branques de la corba que separa l'ordenada que passa per la mitjana aritmètica, la de l'esquerra és més llarga que la de la dreta. El contrari ocorre si $g_1 > 0$ (figura 9).

Cal dir que, així com una distribució simètrica sempre complirà que $g_1 = 0$, l'afirmació contrària no és certa. És a dir, pot haver-hi distribucions no simètriques en què $g_1 = 0$. És per això que el gràfic és, en moltes ocasions, necessari per a decidir sobre l'asimetria d'una distribució de dades.

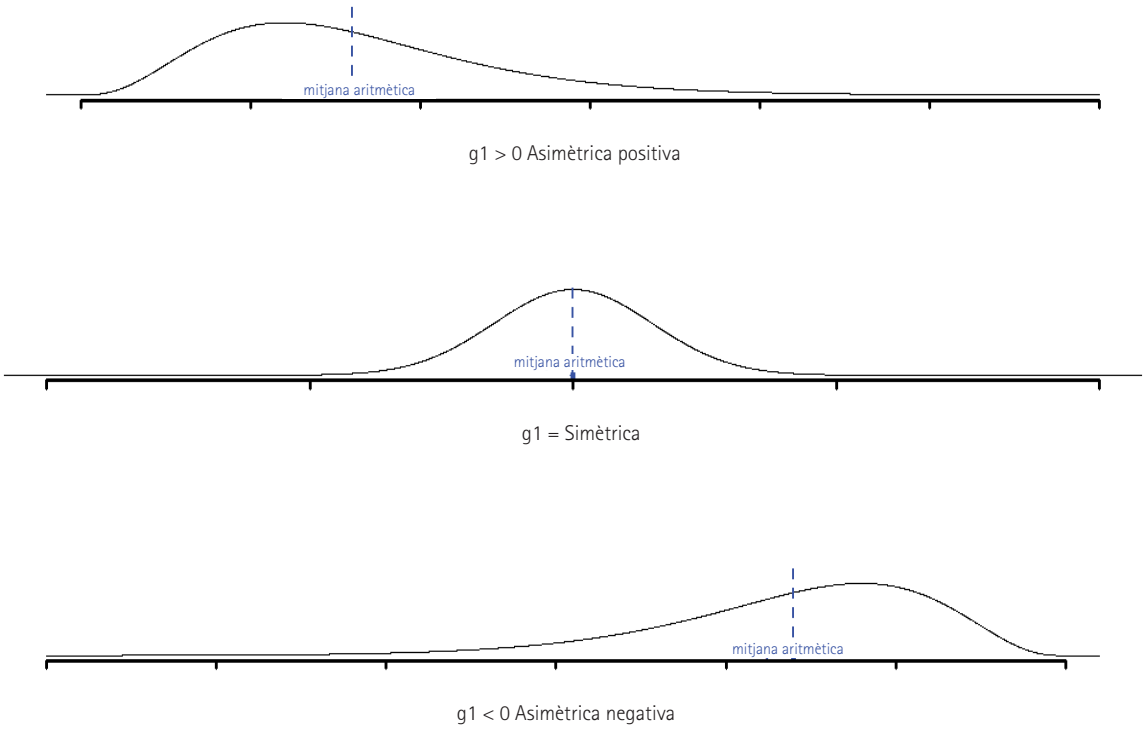


Figura 9

Existeixen altres estadístics per a mesurar l'asimetria, com la proposada per Karl Pearson o el coeficient d'asimetria de Bowley, encara que no es tractaran en aquest text per ser el coeficient d'asimetria de Fisher el més emprat.

Exemple 28

La taula següent mostra l'edat dels alumnes d'una escola que es dedica a ensenyar anglès. Digueu si les dades presenten asimetria.

Edats	Nombre de persones
[5, 10)	2
[10, 15)	9
[15, 20)	8
[20, 25)	3
[25, 30)	3
[30, 35)	5

Per a fer el càlcul, és necessari construir una taula de freqüències. Cal dir que, encara que la mitjana aritmètica i la desviació típica són molt senzilles d'obtenir amb qualsevol calculadora (i fins i tot és recomanable), en aquest exercici es calcularan mitjançant la taula.

Edats	c_i	n_i	$c_i \cdot n_i$	$x_i - \bar{x}$	$(x_i - x)^{-}$	$(x_i - x)^2 \cdot n_i$	$(x_i - x)^3$	$(x_i - x)^3 \cdot n_i$
[5,10)	7,5	2	15	–11,83	140,03	280,06	–1.657,00	–3.313,99
[10,15)	12,5	9	112,5	–6,83	46,69	420,25	–319,08	–2.871,71
[15,20)	17,5	8	140	–1,83	3,36	26,89	–6,16	–49,30
[20,25)	22,5	3	67,5	3,17	10,03	30,08	31,75	95,26
[25,30)	27,5	3	82,5	8,17	66,69	200,08	544,67	1.634,01
[30,35)	32,5	5	162,5	13,17	173,36	866,81	2.282,59	11.412,94
Sumes			580			1.824,17		6.907,22

Per tant:

$$\bar{x} = \frac{\sum_{i=1}^{k}(x_i - \overline{X})\ n_i}{n} = \frac{580}{30} = 19{,}33 \text{ anys}$$

$$s = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \overline{X})^2 n_i}{n}} = \sqrt{\frac{1.824{,}17}{30}} = 7{,}8 \text{ anys}$$

$$m = \frac{\sum_{i=1}^{k}(x_i - \overline{X})^3 n_i}{n} = \frac{6.907{,}22}{30} = 230{,}24 \text{ anys}^3 \text{ i, en conseqüència,}$$

$$g_1 = \frac{m}{s^3} = \frac{230{,}24}{7{,}8^3} = 0{,}49 > 0$$

Per tant, el coeficient d'asimetria de Fisher és positiu i la distribució de dades serà asimètrica positiva. Aquest fet es pot comprovar amb l'histograma (figura 10) de les dades.

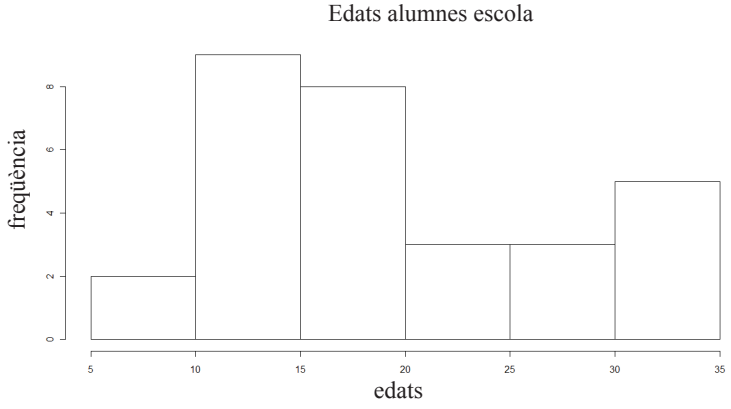


Figura 10

4.4.2. Mesures de curtosi o apuntament

Les mesures de curtosi s'apliquen en distribucions campaniformes, és a dir, en distribucions que tenen una única moda i que són simètriques o lleugerament asimètriques. En essència, les mesures de curtosi tracten d'estudiar la distribució de freqüències en la zona central de la distribució. El grau de concentració de dades al voltant de la mitjana i en la zona central de la distribució donarà com a resultat un histograma més o menys apuntat. Per aquesta raó, les mesures de curtosi també s'anomenen *mesures d'apuntament* o *de concentració central*.

Per a estudiar el grau de curtosi d'una distribució cal prendre un model teòric com a referència, la representació gràfica del qual tinga forma de campana simètrica. No és estrany, doncs, que es prenga el model normal, ja que, com ja s'ha esmentat amb anterioritat, es pot dir que és el model campaniforme per antonomàsia.

D'aquesta manera, prenent aquest model com a referència, es diu que una distribució és leptocúrtica si és més apuntada que la distribució normal. Si és menys apuntada s'anomena *platicúrtica*. Finalment, si té el mateix apuntament que una distribució normal s'anomena *mesocúrtica* (figura 11).

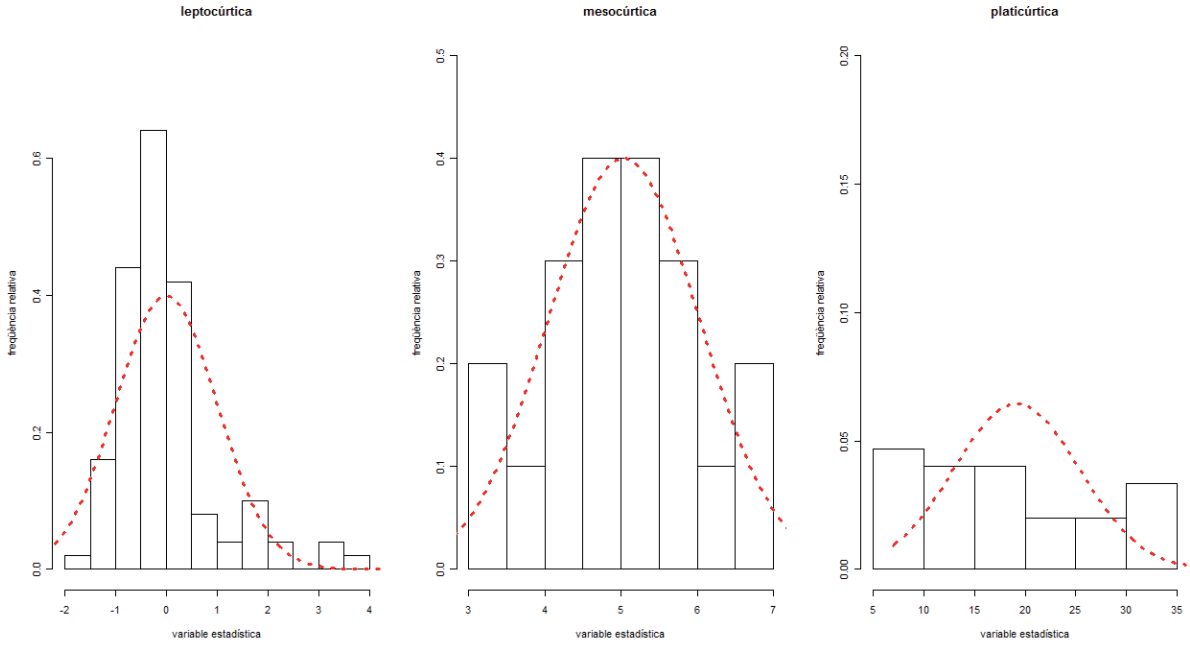


Figura 11

De la mateixa manera que en el cas de l'estudi de l'asimetria, hi ha un coeficient que permet classificar les dades segons la curtosi. En aquest cas, el coeficient no és tan intuïtiu, i per això únicament se'n donarà la definició i la seua interpetació. Com en el cas de l'altra mesura de forma, aquest indicador tampoc té dimensió.

$$g_2 = \frac{\frac{\sum_{i=1}^{k}(x_i - \overline{X})^4 n_i}{n}}{s^4} - 3 = \frac{\frac{\sum_{i=1}^{k}(x_i - \overline{X})^4 n_i}{n}}{\left(\sqrt{\frac{\sum_{i=1}^{k}(x_i - \overline{X})^2 n_i}{n}}\right)^4} - 3, \text{ llavors } g_2 = \frac{\frac{\sum_{i=1}^{k}(x_i - \overline{X})^4 n_i}{n}}{\left(\frac{\sum_{i=1}^{k}(x_i - \overline{X})^2 n_i}{n}\right)^2} - 3$$

La idea de l'apuntament d'una distribució de dades ix de la comparació de la freqüència dels valors centrals d'una distribució amb la freqüència dels valors centrals en un model teòric normal que tinga la mateixa mitjana i la mateixa desviació típica que la distribució que s'està estudiant.

Com que en un model normal es compleix que $\frac{\frac{\sum_{i=1}^{k}(x_i - \overline{X})^4 n_i}{n}}{s^4} = 3$, llavors una distribució serà:

mesocúrtica (normal)	si $g_2 = 0$
leptocúrtica	si $g_2 > 0$
platicúrtica	si $g_2 < 0$

Per a acabar, cal remarcar que l'estudi de la curtosi no implica necessàriament que les distribucions siguen simètriques. Així, per exemple, ens podríem trobar distribucions d'observacions que siguen leptocúrtiques i, al mateix temps, asimètriques positives.

Exemple 29

En aquest exemple s'estudiarà la curtosi per a les mateixes dades que apareixen en l'exemple 28. Per a realitzar el càlcul de l'estadístic g_2 cal construir una taula semblant a la que s'ha construït per al càlcul del coeficient d'asimetria de Fisher.

Així:

Edats	c_i	n_i	$c_i \cdot n_i$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2 \cdot n_i$	$(x_i - \overline{x})^4$	$(x_i - \overline{x})^4 \cdot n_i$
[5, 10)	7,5	2	15	–11,83	140,03	280,06	19.607,78	39.215,56
[10, 15)	12,5	9	112,5	–6,83	46,69	420,25	2.180,37	19.623,34
[15, 20)	17,5	8	140	–1,83	3,36	26,89	11,30	90,38
[20, 25)	22,5	3	67,5	3,17	10,03	30,08	100,56	301,67
[25, 30)	27,5	3	82,5	8,17	66,69	200,08	4.448,15	13.344,45
[30, 35)	32,5	5	162,5	13,17	173,36	866,81	30.054,07	150.270,37
Sumes			580			1.824,17		222.845,76

Segons el que s'ha calculat en l'exemple 28, $\overline{X} = 19{,}33$ anys i $s = 7{,}8$ anys; i per a calcular el coeficient:

$$g_2 = \frac{\sum_{i=1}^{k}(x_i - \overline{X})^4 n_i}{\frac{n}{s^4}} - 3 = \frac{\frac{222845{,}76}{30}}{(7{,}8)^4} - 3, \text{ llavors}$$

$g_2 = 2{,}01 - 3 = -0{,}9 < 0$ i, per tant, aquesta distribució de dades és platicúrtica, és a dir, té un apuntament per sota del model normal.

4.4.3. Diagrama de caixa

Un diagrama de caixa (conegut també com a *box and whisker plot* en anglès), és una representació gràfica de les dades que permet determinar amb molta facilitat i d'una manera visual, la tendència central, la variabilitat, l'asimetria i l'existència de valors anòmals d'un conjunt d'observacions. D'alguna manera, es pot dir que és un dels gràfics que més i millor resumeixen els conjunts de dades.

El diagrama de caixa empra el que David Moore i altres anomenen *el resum dels cinc nombres*: l'observació més petita, l'observació més gran, el primer quartil, la mediana i el tercer quartil. Aquests cinc nombres permeten construir la versió més simple del diagrama de caixa, el qual està format per:

Una caixa (*box*) central que representa les observacions compreses entre el primer i el tercer quartil. Els dos extrems de la caixa són els quartils, i una línia interior i vertical que parteix la caixa en dues parts, correspon a la mediana. És obvi, doncs, que la caixa comprèn el 50% de les observacions (figura 12).

Bigots (*whiskers*). El gràfic es completa en aquesta versió del diagrama, amb dues línies a ambdós costats de la caixa, que uneixen el primer quartil amb l'observació més petita, i el tercer quartil amb l'observació més gran (figura 12).

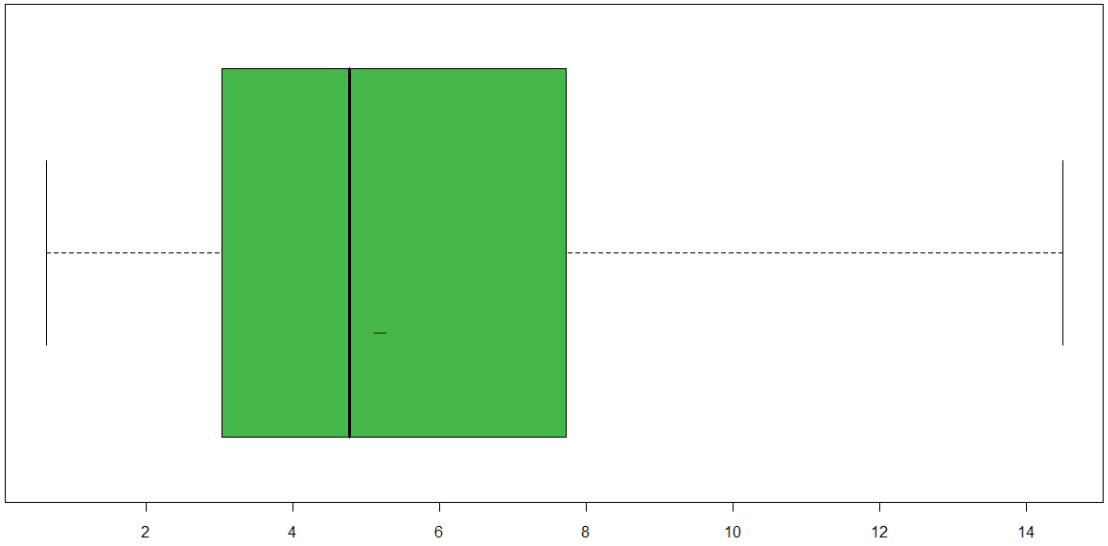


Figura 12

Les dades d'on prové la figura 12 tenen els estadístics següents:

Mín.	1r quartil	Mediana	3r quartil	Màx.,
0,6449	3,5370	5,4770	8,0770	14,4600

i tots s'observen en el gràfic. A més a més, per la forma que tenen els «bigots» i per la ubicació de la mediana dins de la caixa, es pot concloure que l'histograma de les dades no és simètric.

Per altra banda, les diferents versions dels diagrames de caixa tenen en comú que els límits de la caixa són el primer i el tercer quartil, i es diferencien en els valors que determinen els extrems dels bigots. Aquests poden ser els valors més gran i més menut (com en la versió anterior), múltiples del recorregut interquartílic, múltiples de la desviació típica, valors de diferents percentils, etc.

La versió que desenvoluparem és la que va presentar Tukey (2003) en 1977, la qual calcula els límits dels bigots en funció del rang interquartílic. Així doncs, el límit inferior del bigot (costat esquerre) es calcula restant al primer quartil 1,5 vegades el recorregut interquartílic (*Ri*); i el límit superior (costat dret), sumant al tercer quartil 1,5 vegades el *Ri*. El diagrama de caixa de la figura 13 representa les dades següents:

16, 3, 5, 5, 1, 9, 12, 5, 8, 3, 8, 5, 7, 7, 7, 6, 3, 10, 4, 7, 7, 3, 4, 4, 23, 2, 3, 5, 7, 3, 8, 5, 1, 5, 7, 2, 2, 4, 9, 5, 10, 5, 3, 4, 3, 3, 5, 11, 4, 15.

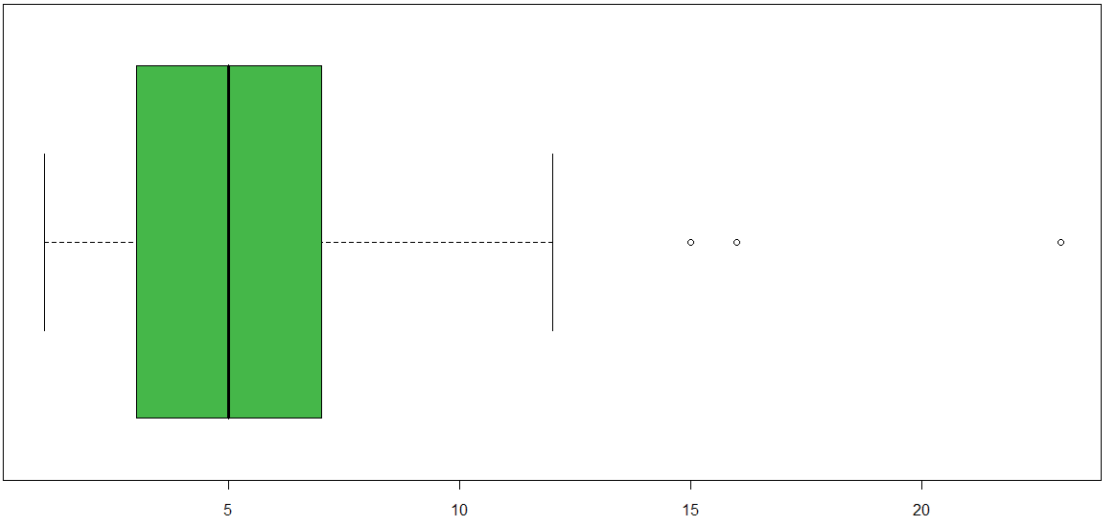


Figura 13

Una vegada representat el gràfic és possible que hàgem observat que hi ha valors que no són dins de la caixa ni dels bigots com ara els següents: 15, 16 i 23. Aquests valors s'allunyen molt de la resta i és possible que aquest fet responga a errors de medició. En conseqüència seria raonable llevar-los de la distribució de dades. Però, quin criteri cal usar? Per a respondre a aquesta qüestió, cal parlar dels valors sospitosos de ser anòmals i dels valors anòmals (o *outliers*, en anglès). Així, un valor es pot considerar sospitós d'*outlier* o d'*anòmal* si no es troba dins dels límits dels bigots però és a menys de tres vegades el recorregut interquartílic del límit del bigot per la dreta o del de l'esquerra. Els valors que s'allunyen més s'anomenen *valors anòmals* o *outliers*.

Els extrems dels bigots representen els valors més gran i més xicotet del total de les observacions que no són considerades anòmales. Els valors sospitosos de ser anòmals, cal tractar-los amb molt de compte, ja que poden ser –o no– errors de medició, i els anòmals poden ser considerats com a no pertanyents a la distribució que s'està considerant.

Pels estadístics que s'empren en la construcció, així com pel gràfic en si, és evident que el diagrama de caixa proporciona una idea de la tendència central de la distribució, així com de la variabilitat. A més a més, també aporta informació sobre les mesures de forma. La proximitat de la mediana als extrems de la caixa i la longitud dels bigots indiquen asimetria en la distribució. Així, si l'asimetria és per la dreta, la mediana estarà prop de l'extrem de la dreta de la caixa i/o la longitud del bigot de la dreta serà més gran que la de l'esquerre. Si la asimetria és per l'esquerra ocorrerà el contrari. Finalment, si la distribució de dades és simètrica, la mediana s'ubicarà al centre de la caixa i els bigots seran simètrics respecte a la mediana (figura 14).

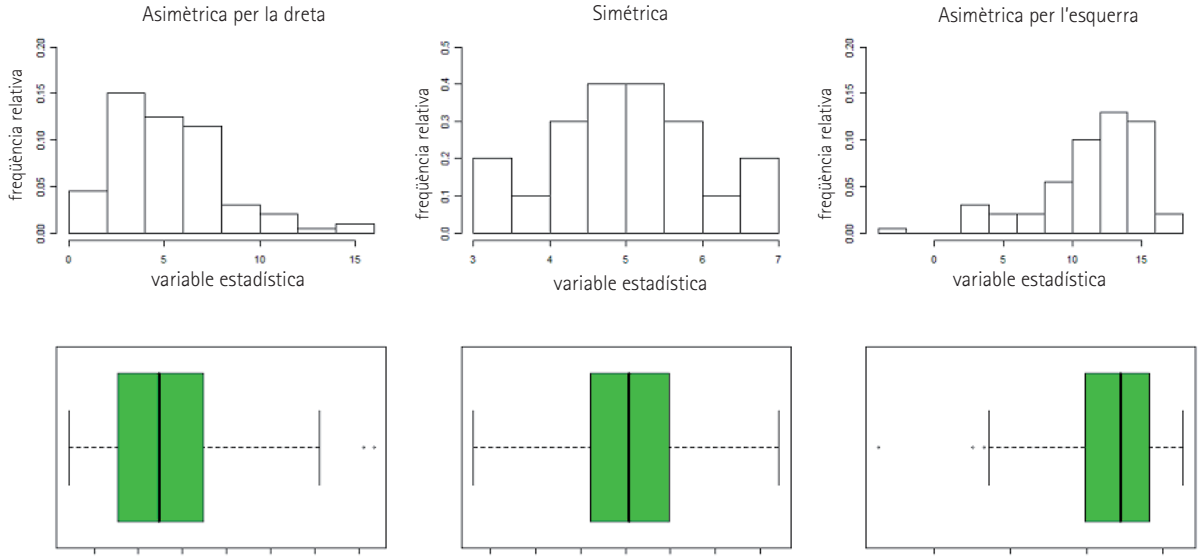


Figura 14

Exemple 30

Les observacions següents corresponen a les reunions setmanals que tenen els treballadors d'una gran multinacional:

5, 5, 6, 6, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 12, 12, 12, 12, 12, 1, 4, 14, 14, 17, 17, 20, 30.

Com que es tracta d'una variable en què hi ha poques dades diferents, no és necessari construir intervals. Així, les dades es poden resumir en una taula com la següent:

Reunions	5	6	8	9	10	12	14	17	20	30
Treballadors	2	2	10	8	6	5	3	2	1	1

Ara cal trobar els estadístics necessaris per a construir el gràfic. Així, realitzant els càlculs de la mateixa manera que a l'epígraf 2, s'obté:

Mín.: 5 Màx.: 30 1r quartil: 8
Mediana: 9 3r quartil: 12

Així doncs, el rang interquartílic és 12 – 8 = 4.

Tot plegat tenim que:

- Els extrems de la caixa són 8 i 12.
- La mediana és 9.

- L'extrem esquerre del bigot és $8 - 1{,}5 \cdot Ri = 8 - 1{,}5 \cdot 4 = 2$. Com que no hi ha cap dada igual a 2, l'extrem del bigot esquerre serà 5.
- L'extrem dret és $12 + 1{,}5 \cdot Ri = 12 + 1{,}5 \cdot 4 = 18$. Com que el valor 18 no és pres per la variable, es considera com a extrem dret del bigot el valor 17. Així doncs, els darrers valors no anòmals són 5 i 17.
- El valor 20 és un valor sospitós de ser anòmal, ja que es troba a més d'$1{,}5 \cdot Ri$ de l'extrem dret del bigot però a menys de $3 \cdot Ri$.
- El valor 30 és un valor anòmal o *outlier*, ja que es troba a més de $3 \cdot Ri$ de l'extrem dret del bigot.

La representació gràfica del diagrama de caixa és el que apareix a la figura següent:

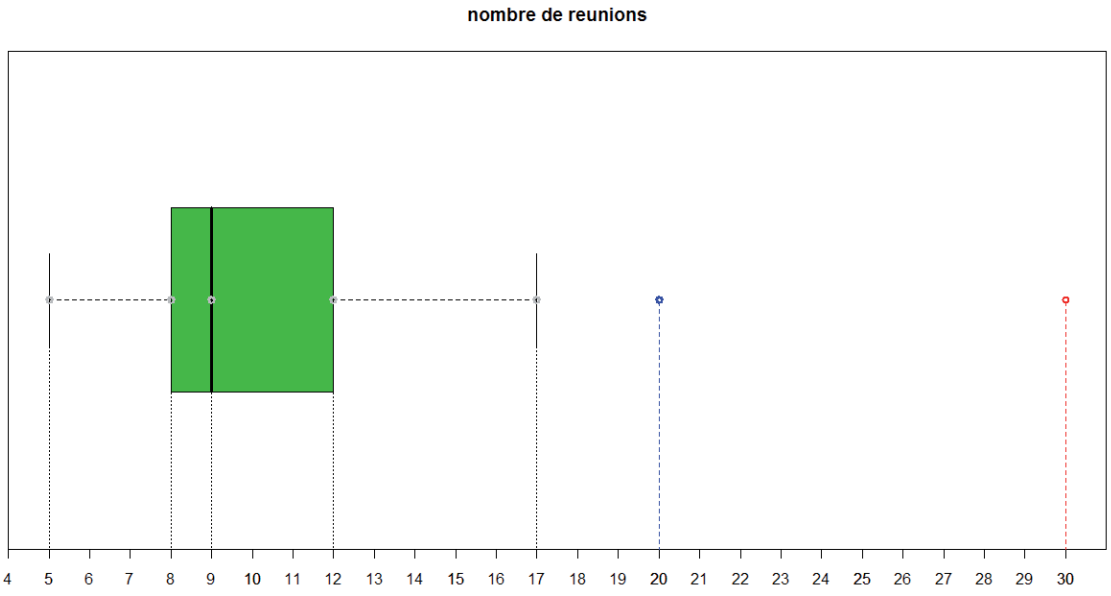


Figura 15

Com es pot observar en el diagrama, la distribució de dades és asimètrica positiva, ja que la mediana està desplaçada cap al costat esquerre i el bigot de la dreta és més llarg que el de l'esquerra.

Nota

Cal observar que el resum dels 5 nombres esmentats en aquest apartat (mínim, primer quartil, mediana, tercer quartil i màxim) i el diagrama de caixa que ix com a conseqüència és, tal com comenta David S. (Moore i altres, 2003), normalment millor que la comparació entre la mitjana i la desviació típica per a descriure distribucions asimètriques o distribucions amb *outliers*. La raó és que en una distribució fortament asimètrica la dispersió és diferent en cadascun dels costats i, en conseqüència, un únic nombre com la desviació típica no descriu la variabilitat de

les dades completament. A més a més, és evident que si en un conjunt d'observacions hi ha *outliers*, aquests afecten notablement tant la mitjana aritmètica com la desviació típica. Per tant, cal emprar la mitjana i la desviació típica per a explicar la dispersió en aquelles distribucions que són lliures d'*outliers* i raonablement simètriques.

Per a acabar, cal comentar que els diagrames de caixa són de molta utilitat a l'hora de comparar diferents distribucions, ja que s'hi poden contrastar ràpidament i d'un colp d'ull, la tendència central i la dispersió. A la figura 16 –extreta d'un exemple del software informàtic lliure per a estadística anomenat *R*– s'observen el nombre d'insectes que han sigut morts per diferents tipus d'insecticides en un experiment real. És evident que la simple observació dels diagrames permet tenir una idea prou clara de l'efectivitat de cadascun dels insecticides.

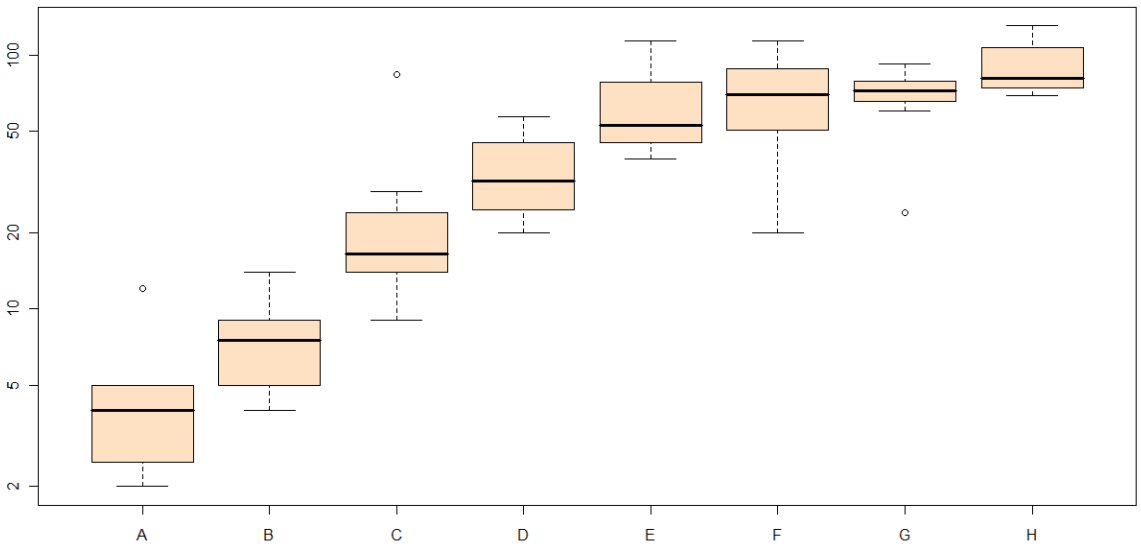


Figura 16

4.5. Mesures de concentració

Les mesures de concentració tracten de mostrar el grau d'igualtat en el repartiment del total dels valors de la variable. Són, per tant, indicadors del grau de distribució de la variable. Amb aquest fi, estan concebuts els estudis sobre concentració. Així, per exemple, donada la taula 1, que mostra els sous que una empresa paga als treballadors, la pregunta que ajuda a contestar aquest tipus de mesures és si la nòmina total que paga l'empresa està ben repartida entre els treballadors o no. Existeixen diferents índexs que donen resposta a aquesta qüestió, com ara l'índex geomètric o raó de concentració, l'índex de Theil o l'índex de Gini. Al text es tractaran l'índex de Gini i la corba de Lorenz.

$[L_{i-1}, L_i)$ (en desenes)	n_i	$[L_{i-1}, L_i)$ (en desenes)	n_i
[0, 50)	23	[250, 300)	8
[50, 100)	72	[300, 350)	14
[100, 150)	62	[350, 400)	7
[150, 200)	48	[400, 450)	5
[200, 250)	19	[450, 500)	2

Taula 1

Es denomina *concentració* el grau d'equitat en el repartiment de la suma total dels valors de la variable considerada (renda, salaris, etc.). En el cas que estem considerant, l'empresa repartiria un total de 388.500 € entre els 260 treballadors.

Les infinites possibilitats que poden adoptar els valors, es troben entre els dos extrems:

Concentració màxima. Quan només un percep el total i els altres res; en aquest cas, estem davant d'un repartiment no equitatiu:

$$\textit{el que rep } x_1 = \textit{el que rep } x_2 = \ldots = \textit{el que rep } x_{k-1} = 0 \quad i \quad \textit{el que rep } x_k = \textit{el total}$$

Concentració mínima. Quan el conjunt total de valors de la variable està repartit per igual, en aquest cas estem davant d'un repartiment equitatiu:

$$\textit{el que rep } x_1 = \textit{el que rep } x_2 = \ldots = \textit{el que rep } x_{k-1} = \textit{el que rep } x_k$$

Hi ha diferents mesures de concentració, però en el text s'estudiarà l'índex de Gini; per ser un coeficient, serà un valor numèric. Per a obtenir-lo és necessari realitzar un conjunt de càlculs.

Se suposa que es té una distribució de rendes ($x_i \cdot n_i$), on i pren els valors de *1* fins a k (per exemple, x_i són els sous i n_i el nombre de persones que cobren aquest sou), de la qual es formarà una taula amb les columnes següents:

1) Els productes $x_i \cdot n_i$ indicaran la renda total percebuda pels n_i rendistes de renda individual x_i.
2) Les freqüències absolutes acumulades N_i.
3) Els totals acumulats u_i, que es calculen de la forma següent:

$$
\begin{aligned}
u_1 &= x_1 n_1 \\
u_2 &= x_1 n_1 + x_2 n_2 \\
u_3 &= x_1 n_1 + x_2 n_2 + x_3 n_3 \\
u_4 &= x_1 n_1 + x_2 n_2 + x_3 n_3 + x_4 n_4 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
u_k &= x_1 n_1 + x_2 n_2 + x_3 n_3 + x_4 n_4 + \cdots\cdots\cdots + x_k n_k
\end{aligned}
$$

Per tant, es pot dir que:

$$u_j = \sum_{i=1}^{j} x_i \cdot n_i \text{ per a qualsevol valor de } j \text{ des de 1 fins a } k.$$

4) La columna total de freqüències acumulades relatives, que s'expressa en tant per cent i que es representa per p_i, estarà donada per la notació següent:

$$p_i = \frac{N_i}{n}.$$

5) La columna de renda acumulada relativa, que s'expressa en tant per cent i que es representa per l'expressió:

$$q_i = \frac{u_i}{u_k}.$$

Per tant, ja es pot confeccionar la taula:

x_i	n_i	$x_i n_i$	N_i	u_i	$p_i = \frac{N_i}{n}$	$q_i = \frac{u_i}{u_k}$	$p_i - q_i$
x_1	n_1	$x_1 n_1$	N_1	u_1	p_1	q_1	$p_1 - q_1$
x_2	n_2	$x_2 n_2$	N_2	u_2	p_2	q_2	$p_2 - q_2$
...	...	...	...	...	...	...	...
x_k	n_k	$x_k n_k$	N_k	u_k	100	100	0

Com es pot veure, l'última columna és la diferència entre les dues penúltimes; aquesta diferència seria 0 per a la concentració mínima, en la qual es compleix $p_i = q_i$ per a qualsevol i, per tant la diferència seria 0.

Analíticament l'índex de Gini:

$$I_G = \frac{\sum_{j=1}^{k-1}(p_i - q_i)}{\sum_{j=1}^{k-1} p_i}.$$

Aquest índex prendrà els valors:

- $i_G = 0$ quan $p_i = q_i$ concentració mínima
- $i_G = 1$ quan $q_i = 0$ concentració màxima

Per altra banda, si es representen gràficament els q_i a l'eix vertical i els p_i a l'horitzontal s'obtindrà la corba de concentració o corba de Lorenz. Es pot comprovar que aquesta corba resultant sempre apareixerà per sota de la diagonal del primer quadrant, la qual representa la concentració mínima. A més a més, com més s'aproxime aquesta corba a la diagonal, més petita serà la concentració. La figura 17 mostra la interpretació de diferents corbes de Lorenz.

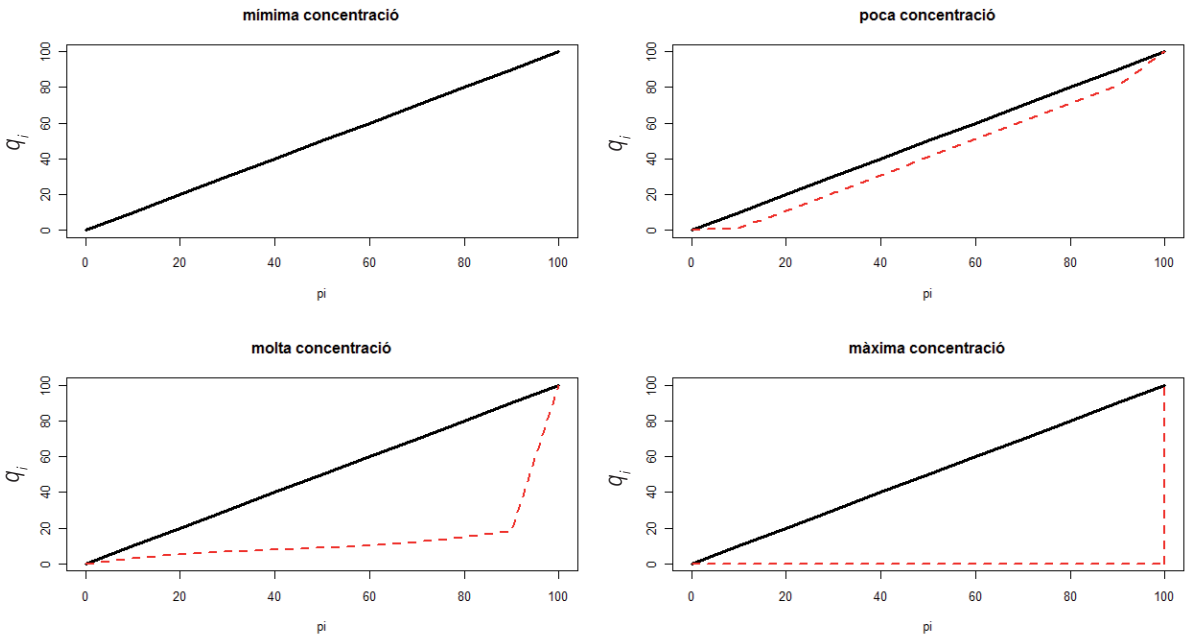


Figura 17

Exemple 31

La taula següent mostra en centenes d'euro el sou que cobren els treballadors d'una determinada empresa. Digues si la nòmina total que reparteix l'empresa és equitativa o si, al contrari, està molt concentrada en poques persones.

$[L_{i-1}, L_i)$	n_i
[0, 50)	23
[50, 100)	72
[100, 150)	62
[150, 200)	48
[200, 250)	19
[250, 300)	8
[300, 350)	14
[350, 400)	7
[400, 450)	5
[450, 500)	2

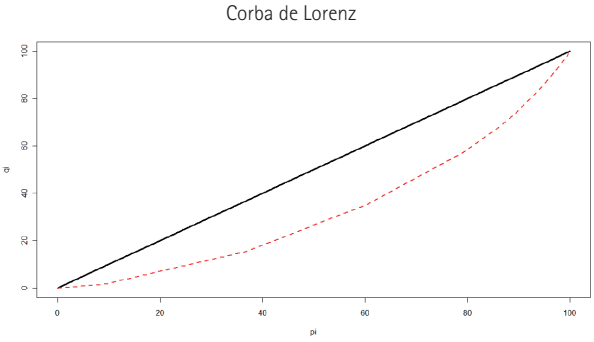
Per a respondre a la pregunta cal calcular l'índex de Gini o la corba de Lorenz. Per a fer-ho, és convenient construir una taula en què apareguen els p_i, q_i i la seua diferència. Així doncs:

$[L_{i-1}, L_i)$	c_i	n_i	N_i	$X_i n_i$	u_i	$q_i = (u_i/u_k)\ 100$	$p_i = (N_i/n)\ 100$	$p_i - q_i$
[0, 50)	25	23	23	575	575	1,48	8,85	7,37
[50, 100)	75	72	95	5.400	5975	15,38	36,54	21,16
[100, 150)	125	62	157	7.750	13.725	35,33	60,38	25,06
[150, 200)	175	48	205	8.400	22.125	56,95	78,85	21,90
[200, 250)	225	19	224	4.275	26.400	67,95	86,15	18,20
[250, 300)	275	8	232	2.200	28.600	73,62	89,23	15,61
[300, 350)	325	14	246	4.550	33.150	85,33	94,62	9,29
[350, 400)	375	7	253	2.625	35.775	92,08	97,31	5,22
[400, 450)	425	5	258	2.125	37.900	97,55	99,23	1,68
[450, 500)	475	2	260	950	38.850	100,00	100,00	0,00
Sumes		260		38.850			651,15	125,48

Índex de concentració de Gini

U_k

$$I_G = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = \frac{125{,}48}{651{,}15} = 0{,}193$$



Com es pot comprovar, tant l'índex de Gini (0,193) com la corba de Lorenz indiquen que la concentració no és elevada. En conseqüència es pot concloure que la nòmina està repartida prou equitativament.

4.6. Problemes proposats

En aquest epígraf es plantejaran un conjunt de problemes per a la resolució dels quals és necessari conèixer la teoria desenvolupada al llarg de la unitat.

Exercici 1

Si les puntuacions obtingudes en les proves d'accés a un lloc de treball han estat les que es mostren en la taula següent:

Puntuacions	Nombre de persones
[0, 20)	26
[20, 40)	25
[40, 60)	55
[60, 80)	45
[80, 100)	25

a) Calcula quina variable s'està estudiant i de quina població.
b) Calcula la nota mitjana de les puntuacions obtingudes.
c) Representa gràficament la distribució de dades i marca la mitjana.

Exercici 2

Els salaris anuals de 4 individus són de 150.000, 160.000, 165.000 i 200.000 €. Calcula'n el salari mitjà.

Ara entra a treballar una nova persona en l'empresa, i percep un salari de 500.000 €. Es veurà afectat el salari mitjà després d'aquesta incorporació? Creus que la mitjana és una mesura de centralització adequada en tots dos casos? En cas que no ho siga, proposa i calcula una altra mesura de centralització més adequada.

Exercici 3

Ja fa 5 anys que una persona té un fons d'inversió, els interessos que li ha estat rendint el fons al llarg dels cinc anys (2000-2004) han sigut: 4,25; 3,75; 2,75; 2,25 i 2,45. Quin ha sigut l'interès mitjà?

Exercici 4

Un inversor de borsa ha comprat accions d'una mateixa empresa en cinc ocasions: els preus respectius per acció han sigut de 20,48; 23,04; 19,20; 25,60 i 30,72 euros. Calculeu el preu mitjà per acció en els supòsits següents:

a) Les cinc vegades va adquirir el mateix nombre d'accions *N*.
b) En totes les compres va emprar la mateixa quantitat de diners *C*.

Exercici 5

En una enquesta sobre el nombre de llocs de treball al llarg de la vida de 100 persones, s'obtingueren els resultats següents:

Llocs de treball	1	3	5	6	10	11	18
Persones	20	20	20	15	15	8	2

Calcula'n la mitjana, la mediana, la moda, el tercer quartil, el cinquè decil i el 76è percentil i interpreta'ls.

Exercici 6

La distribució d'edats del cens electoral de residents l'1 de gener de 1999 per a les comunitats autònomes d'Aragó i Canàries, en tant per cent, és la següent:

Edats	Aragó	Canàries
16-18	3,55	4,35
18-30	21,56	29,99
30-50	31,63	35,21
50-70	28,14	21,97
70-90	15,12	8,48

a) Representa sobre els mateixos eixos de coordenades els histogrames de la distribució de l'edat per a les dues comunitats autònomes (empra-hi traç o colors distints). Quines conclusions obtens a la vista dels histogrames?
b) Calcula l'edat mitjana per a les dues comunitats. Compara-les. Què indiquen aquests resultats?
c) En quina comunitat són més disperses les observacions?

Exercici 7

En la taula següent estan reflectides les ajudes europees, en milers de pessetes, per al Fons Europeu de l'any 1995:

Import de l'ajuda	Nombre de projectes
0-100	10
100-250	15
250-500	20
500-1000	15

a) Calcula l'ajuda mitjana i digues si el valor és o no representatiu del conjunt de dades.
b) Calcula l'ajuda màxima concedida al 60% dels projectes menys afavorits en el repartiment.
c) Si l'any següent les ajudes augmenten un 5% sobre el valor inicial, i es manté el criteri de repartiment, quina serà l'ajuda mitjana ara? Continua sent representativa?
d) Està repartit equitativament el total de les ajudes?

TEMA 5

Distribució de dues variables estadístiques. Regressió lineal

Objectius tema 5

D'una distribució de dades bidimensional,
- saber analitzar i extraure informació;
- saber extraure conclusions de l'anàlisi tant de les distribucions marginals com de les condicionades;
- distingir gràficament i analíticament si les dues variables tenen relació lineal;
- construir la recta de regressió lineal d'una variable estadística respecte de l'altra;
- saber predir el valor d'una variable a partir d'un valor de l'altra mitjançant la recta de regressió. Conèixer la fiabilitat de les prediccions.

5.1. Introducció

Normalment, en qualsevol investigació no s'estudia una única variable dels individus que formen la mostra, sinó que en moltes ocasions se n'estudien més. Així, si es desitja estudiar el rendiment dels treballadors d'una empresa, pot ser útil conèixer de cadascun: l'edat, el sou, el nivell d'estudis, les hores que treballa, el nombre de persones que té al seu càrrec, etc. És a dir, per a cada individu de la mostra s'obté un vector o registre en què cada component és el valor d'una de les variables subjectes a estudi; en l'exemple que s'està considerant associarem un vector a cada individu (35 anys, 24.500 €, diplomat, 47 hores setmanals, 2 persones al seu càrrec...).

Aquest fet origina que l'investigador es plantege, a més a més de l'estudi individualitzat de cadascuna de les variables, l'estudi conjunt de totes o d'algunes d'aquestes. D'aquesta manera és possible conèixer si existeix algun tipus de relació funcional o estadística entre les variables. Així, les observacions poden manifestar que aquelles persones amb més titulació tenen més individus al seu càrrec, o que a mesura que va augmentant l'edat dels treballadors també ho fa el sou. A més a més, si aquesta relació existeix, tal vegada es pot trobar una fórmula matemàtica que relacione formalment les variables.

D'altra banda, la nomenclatura canvia si s'estudien conjuntament diferents variables. Així, si es realitza l'estudi de dues variables es diu que es treballa amb variables bidimensionals; si en són tres, amb variables tridimensionals; i si en són més de tres, amb variables pluridimensionals.

En general, l'estudi conjunt de diverses variables és molt ampli. En la unitat s'estudiaran les variables bidimensionals. Concretament, es mostraran les diferents maneres de resumir les observacions utilitzant les taules de contingència i el diagrama de dispersió, i s'analitzaran les variables marginals –individualment– i les condicionades (valors que pren una de les variables quan l'altra en pren un de fix). Posteriorment, es presentaran els procediments estadístics que permeten saber quin grau de relació lineal hi ha entre dues variables, construir una funció lineal que les relacione i realitzar prediccions del valor d'una de les variables una vegada fixat un valor de l'altra.

5.2. Distribucions estadístiques bidimensionals: taules i gràfics

Quan es volen estudiar dues característiques observables sobre una mateixa mostra o població, cada una de les variables que constitueixen la variable bidimensional (X, Y) es denomina *component* o *variable marginal* d'aquesta, i pot ser tant un atribut com una variable quantitativa. En qualsevol cas, en realitzar-se el treball de recollida de dades s'obté un conjunt de parells ordenats del tipus:

$$\{(x_1, y_1), (x_1, y_1), ..., (x_2, y_1), (x_2, y_1), ..., (x_2, y_1), (x_2, y_2), ..., (x_1, y_j), ..., (x_1, y_j), ..., (x_h, y_k), ..., (x_h, y_k)\}$$

Per exemple, si es considerara X la variable «dies d'estudi per a un examen d'estadística» i Y la variable «nota obtinguda per a un conjunt d'estudiants», les dades recollides serien del tipus:

$$\{(5,3); (6,5); (5,3); (6,5); (5,7)\}$$

En les dades, cada observació es repeteix un nombre de vegades determinat. Així, una primera manera de representar el conjunt de dades és mitjançant la terna $\{(x_i, y_j), n_{i,j}\}$ en què:

- x_i representa els valors de la variable X,
- y_i representa els valors de la variable Y,
- $n_{i,j}$ és el nombre de vegades que es repeteix la dada (x_1, y_j), és a dir, la seua freqüència absoluta.

Seguint l'exemple tindríem que:

$x_1 = 5$	$y_1 = 3$	$n_{11} = 2$	
$x_1 = 5$	$y_3 = 7$	$n_{13} = 1$	
$x_2 = 6$	$y_2 = 5$	$n_{22} = 2$.	La resta de $n_{i,j} = 0$.

Per altra part, és evident que tenir tres-cents parells ordenats d'observacions aclareix ben poc la informació. No és possible observar-hi quasi res. En conseqüència, és necessari representar les dades de manera que siguen més comprensibles i en faciliten l'estudi.

Nota

Si es denota per n el nombre total de dades, aleshores es compleix la relació següent:

$$n_{11} + n_{12} + n_{13} + ... + n_{1k} + n_{21} + n_{22} + ... + n_{2k} + ... + n_{h1} + ... + n_{hk} = n.$$

Utilitzant notació matemàtica, es té que: $\sum_{i=1}^{h}\sum_{j=1}^{k} n_{ij} = n$.

En l'exemple que s'està considerant, $h = 2$, $k = 3$ i és clar que $n_{11} + n_{13} + n_{22} = 5$.

Nota

D'ara endavant, es considerarà una distribució bidimensinal com la presentada anteriorment: $\{(x_i, y_j), n_{i,j}\}$ en què hi ha h valors diferents de la variable X i k valors diferents de la variable Y. Es considerarà que en total són n observacions bivariants.

Tenint en compte l'anterior, es pot obtenir la primera forma de recollir les dades: mitjançant una taula amb tres columnes (taula 1). La primera correspon als diferents valors que pren la variable X, la segona als de la variable Y, i la darrera a la freqüència absoluta de cada observació (x_1, y_j). A més, convé ordenar almenys una de les dues variables en sentit creixent. Així:

X	Y	n_{ij}
x_1	y_1	n_{11}
x_1	y_2	n_{12}
.	.	.
.	.	.
x_i	y_j	n_{ij}
.	.	.
.	.	.
x_h	y_k	n_{hk}

Taula 1

Aquesta taula presenta i ordena les dades; no obstant això, en algunes ocasions no és la taula més adequada.

Encara que les dades estigueren agrupades en intervals, la representació mitjançant aquesta taula es realitzaria de forma similar. En ocasions s'utilitzaria la marca de classe com a representació de l'interval.

Exemple 1

L'any 1999 els residents d'un petit poble estaven preocupats per l'increment del cost de l'habitatge a la zona. L'alcalde considerava que els preus de l'habitatge fluctuaven amb els preus dels solars. Els costos (en milers d'euros) de les vivendes i dels terrenys sobre els quals es van construir les cases són els següents:

Valor del terreny	Cost de l'habitatge
7	67
7	67,15
7	67
6,9	63
6,9	63
5,5	60
3,7	54
3,7	54
5,9	58
3,8	36
3,8	36
3,8	36
8,9	76
8,9	76
9,6	87
9,9	89
9,6	87
9,9	89
10	92
10	92
5,9	58
3,8	36
9,6	87
9,6	87
8,9	76
3,7	54
5,5	60
3,8	36
8,9	76
9,9	89

Com es pot apreciar, les dades recollides en la taula anterior aporten poca informació. Construirem, ja que no hi ha molts parells diferents, la taula amb les tres columnes. Se suposarà que:

X = valor del terreny,

Y = valor de l'habitatge.

X	Y	n_j
3,7	54	3
3,8	36	4
5,5	60	2
5,9	58	2
6,9	63	2
7	67	2
7	67,15	1
8,9	76	4
9,6	87	4
9,9	89	3
10	92	2

5.2.1. Taules de doble entrada o de contingència

La taula anterior, tal com s'ha comentat abans, algunes vegades és incòmoda. Per això és preferible utilitzar la taula de doble entrada, que permet extraure molta més informació de la distribució de dades. La taula 2 presenta forma de rectangle, tal com s'observa tot seguit:

X \ Y	y_1	y_2	. . .	y_j	. . .	y_k	$n_{i\cdot}$
x_1	n_{11}	n_{12}	. . .	n_{1j}	. . .	n_{1k}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	. . .	n_{2j}	. . .	n_{2k}	$n_{2\cdot}$
. . .	. . .	. . .	. . .	. .	. . .	. . .	. . .
x_i	n_{i1}	n_{i2}	. . .	n_{ij}	. . .	n_{ik}	$n_{i\cdot}$
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
x_h	n_{h1}	n_{h2}	. . .	n_{hj}	. . .	n_{hk}	$n_{h\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	. . .	$n_{\cdot j}$	. . .	$n_{\cdot k}$	n

Taula 2

En la primera fila se situen les diferents categories o valors que pren una de les components, i en la primera columna els valors o les categories relatives a la segona. D'aquesta forma, qualsevol nombre que apareix en una cel·la interior de la taula de doble entrada és la freqüència absoluta conjunta de la dada bivariant formada pels valors corresponents ubicats a la primera fila i a la primera columna. És a dir; n_{ij} és la freqüència absoluta conjunta de la dada (x_i, y_j). En algunes ocasions també se sol representar en cada cel·la la freqüència relativa conjunta, a més a més de l'absoluta. A més a més, les variables numèriques convé ordenarles en sentit creixent.

D'altra banda, els valors que apareixen a l'última columna i l'última fila corresponen a les freqüències absolutes dels valors de les variables de la primera columna i la primera fila, respectivament. És a dir, $n_{i\cdot}$ representa la freqüència absoluta del valor x_i de la variable X; és a dir, x_i apareix $n_{i\cdot}$ vegades a la distribució de dades.

Si les dades estigueren agrupades en intervals, la representació mitjançant aquesta taula es realitzaria de forma similar, utilitzant la marca de classe com a representació de l'interval.

La taula de doble entrada s'anomena *taula de contingència* o *taula de correlació*, segons si alguna de les components és un atribut o totes dues són variables quantitatives.

Exemple 2

Amb les mateixes dades que en l'exemple anterior, la taula de doble entrada queda:

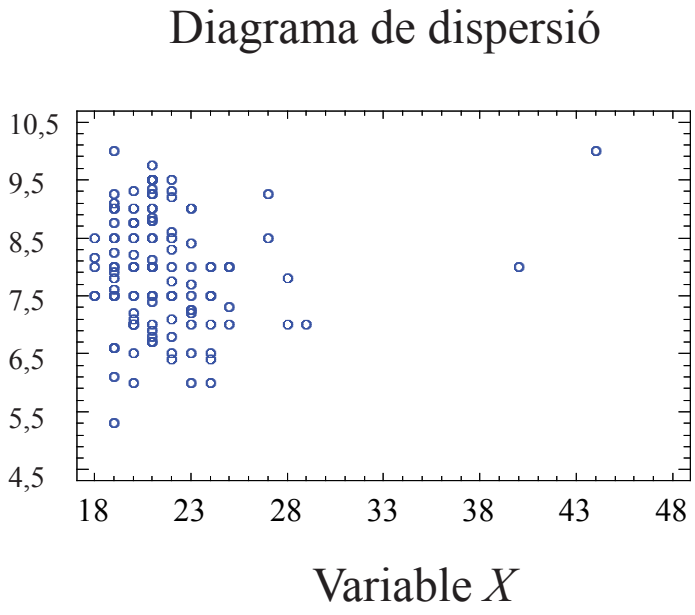
X \ Y	36	54	58	60	63	67	67,15	76	87	89	92	n_i
3,7		3										3
3,8	4											4
5,5				2								2
5,9			2									2
6,9					2							2
7						2	1					3
8,9								4				4
9,6									4			4
9,9										3		3
10											2	2
n_j	4	3	2	2	2	2	1	4	4	3	2	29

Les cel·les buides representen unes freqüències absolutes conjuntes iguals a 0.

De la mateixa manera que ocorria amb les distribucions de dades unidimensionals, les representacions gràfiques faciliten la comprensió de la distribució amb tan sols un colp d'ull.

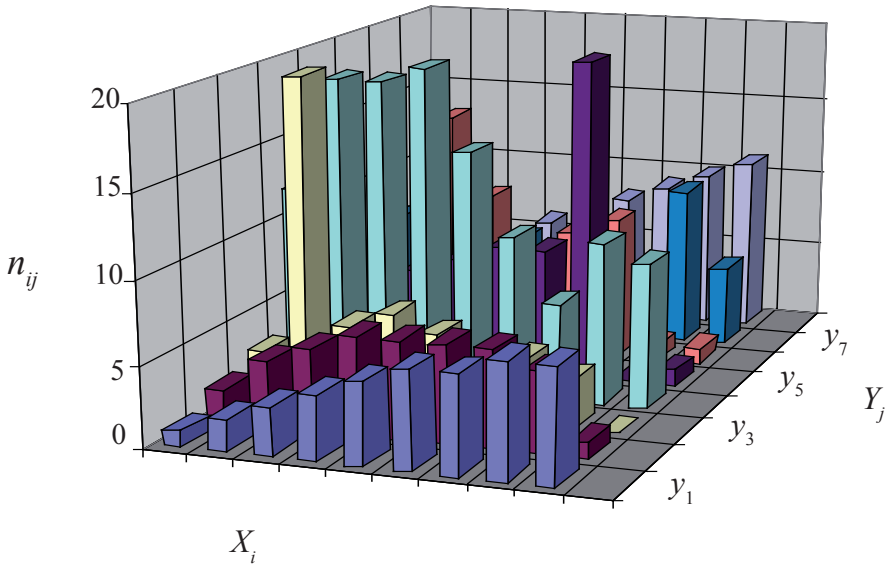
5.2.2. Representacions gràfiques: diagrama de dispersió o núvol de punts

La representació gràfica de la distribució de freqüències d'una variable bidimensional (X, Y) varia sensiblement segons la naturalesa de les variables. Si aquestes són discretes, la representació comuna de la distribució conjunta és el núvol de punts o diagrama de dispersió, el qual es construeix situant sobre l'eix horitzontal d'un sistema cartesià els diferents valors de la variable X; sobre el vertical els de la variable Y; i un punt en la posició (x_i, y_i) si és que aquesta observació té una freqüència absoluta conjunta d'un punt. Si en té més d'un punt, hi ha diferents possibilitats per a representar-ho: dibuixant punts de diferents superfícies (la qual representarà la freqüència), escrivint la freqüència al costat del punt marcat, etc.



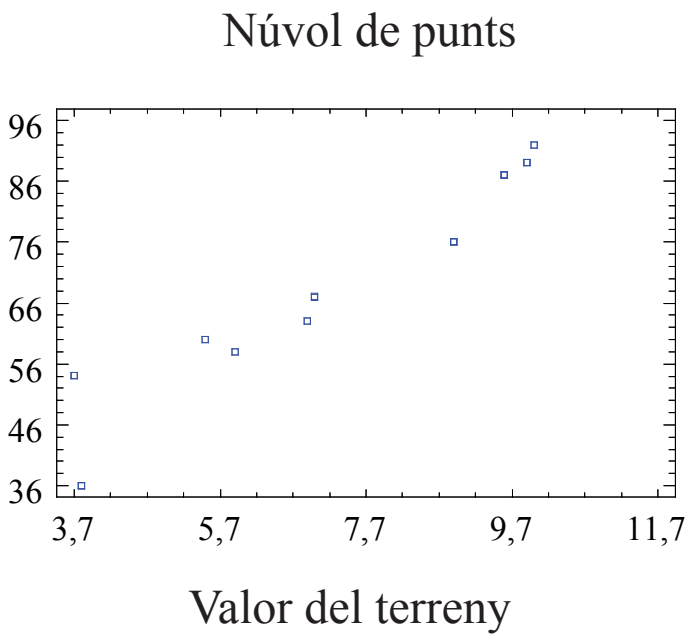
Si alguna de les dades està agrupada en intervals, es pot tenir el criteri de representar les dades com si les marques de classe dels intervals foren els valors de les variables.

Una altra manera de representar les dades és emprant un gràfic tridimensional com el següent:



Exemple 3

En l'exemple 1 que s'està considerant relatiu al valor del terreny i el cost de l'habitatge, el núvol de punts és:



5.3. Distribucions estadístiques marginals i condicionades

És evident que la taula de doble entrada esmentada a l'epígraf anterior ofereix molta informació. De fet, és possible analitzar-ne separadament cada variable component de la variable conjunta, així com una variable condicionada a un valor concret de l'altra.

5.3.1. Distribucions marginals

Establerta una distribució conjunta de freqüències d'una variable bidimensional (X, Y) mitjançant la corresponent taula de doble entrada, és senzill obtenir les distribucions de les variables unidimensionals X i Y que la componen. Aquestes distribucions s'anomenen *distribucions marginals*.

Si les variables X i Y són no agrupades o qualitatives, la distribució marginal de X s'obté de la taula de doble entrada, adjuntant a cada un dels valors $x_1, x_2, ..., x_h$ de la variable estadística X, les seues freqüències absolutes, que es donen a l'última columna de la taula. Així mateix, s'obté la distribució marginal de Y. En aquest cas els valors de la variable $y_1, y_2, ..., y_k$ i les seues freqüències absolutes apareixen en la primera i l'última fila, respectivament. Si les variables estigueren agrupades en intervals, es realitzaria el mateix procediment prenent la marca de classe com a representant de l'interval i, per tant, com a valor de la variable estadística. Així doncs, prenent com a referència la taula de doble entrada de l'epígraf anterior, les distribucions marginals són:

Distribució marginal de X

X	n_i
x_1	n_1
x_2	n_2
.	.
.	.
x_i	n_i
.	.
.	.
x_h	n_h
	n

Distribució marginal de Y

Y	n_j
y_1	n_1
y_2	n_2
..	.
.	.
y_j	n_j
.	.
.	.
y_k	n_k
	n

Cal dir que cada distribució marginal pot ser tractada estadísticament com una variable unidimensional: es poden calcular les mesures de posició i de dispersió de la mateixa manera que s'ha fet a la unitat 4. Així, per exemple, les mitjanes i les variàncies de cada variable serien:

$$\overline{X} = \frac{\sum_{i=1}^{h} n_{i\cdot} \cdot x_i}{n} \qquad \overline{Y} = \frac{\sum_{j=1}^{k} n_{\cdot j} \cdot y_j}{n}$$

$$S_X^2 = \frac{\sum_{i=1}^{h} n_{i\cdot} (x_i - \overline{X})^2}{n} \qquad S_Y^2 = \frac{\sum_{j=1}^{k} n_{\cdot j} \cdot (y_j - \overline{Y})^2}{n}$$

Nota

Cal remarcar la igualtat $\sum_{i=1}^{h} n_{i\cdot} = \sum_{j=1}^{k} n_{\cdot j} = \sum_{i=1}^{h} \sum_{j=1}^{k} n_{ij} = n.$

Exemple 4

En l'exemple 1 que s'està considerant del valor del terreny i el cost de l'habitatge:

Distribució marginal: valor del terreny			**Distribució marginal: cost de l'habitatge**	
X	n_i		Y	n_j
3,7	3		36	4
3,8	4		54	3
5,5	2		58	2
5,9	2		60	2
6,9	2		63	2
7	3		67	2
8,9	4		67,15	1
9,6	4		76	4
9,9	3		87	4
10	2		89	3
	29		92	2
				29

5.3.2. Distribucions condicionades

De la taula de doble entrada, també és possible obtenir, a més de les distribucions marginals, unes altres distribucions. Si s'associen als valors de Y les freqüències corresponents a la fila en què està ubicat el valor x_i de X, resulta la distribució condicionada de Y a x_i (distribució de la variable $Y/X = x_i$). Anàlogament, però tenint present les columnes enlloc de les files s'obtindria la distribució de X condicionada a y_j de Y (distribució de la variable $X/Y = y_j$).

Cal, però, observar que la suma de les freqüències de X condicionada al valor y_j coincideix amb la freqüència marginal d'aquest valor, i de la mateixa manera, la suma de les freqüències de Y condicionada a x_i és igual a la freqüència marginal de x_i.

Així, prenent la taula de doble entrada anterior, les distribucions condicionades són:

Distribució condicionada $X/Y = y_j$

x/y_j	n_{ij}
x_1	n_{1j}
x_2	n_{2j}
...	...
x_i	n_{ij}
...	...
x_h	n_{hj}
	n_j

Distribució condicionada $Y/X = x_i$

Y/x_i	n_{ij}
y_1	n_{i1}
y_2	n_{i2}
...	...
y_j	n_{ij}
...	...
y_k	n_{ik}
	n_i

Les distribucions condicionades són susceptibles del mateix resum quantitatiu que el que es du a terme amb qualsevol variable unidimensional.

Exemple 5

Si en l'exemple 1 es vol conèixer la distribució del preu de l'habitatge quan el preu del solar és de 7.000 euros, la distribució condicionada és *Preu habitatge / Preu solar* = 7.000 i es pot obtenir de la taula de doble entrada següent:

X \ Y	36	54	58	60	63	67	67,15	76	87	89	92	n_i
3,7		3										3
3,8	4											4
5,5				2								2
5,9			2									2
6,9					2							2
7						2	1					3
8,9								4				4
9,6									4			4
9,9										3		3
10											2	2
n_j	4	3	2	2	2	2	1	4	4	3	2	29

i de la taula associada:

Distribució condicionada: preu solar 7.000 €	
***Y/X* = 7**	n_j
67	2
67,15	1
	3

5.4. Correlació lineal

Quan s'estudien dues variables estadístiques conjuntament, és important saber si tenen algun tipus de relació. Així, si es recolliren tres-centes dades en què la primera variable fóra l'alçada d'una persona i la segona, el resultat de llançar un dau, segurament la intuïció diria que entre ambdues variables no hi ha cap tipus de relació. Si, per contra, es consideraren les variables «hores extra que treballa una persona» i «sou que cobra mensualment», la relació canviaria, fins al punt de conèixer el sou d'un individu si se saben les hores extres que fa. Es podria dir que totes dues variables estan lligades per una relació funcional. Tanmateix, si es consideraren les variables «hores de preparació d'un examen» i «nota obtinguda», la intuïció establiria que sí que hi ha alguna relació entre totes dues variables, molt més forta que en el primer cas, però més dèbil que en el segon.

Resumint:

CASOS	**VARIABLES**	**FUNCIÓ QUE LES RELACIONA**	**TIPUS DE RELACIÓ**
Cas 1	X = alçada Y = resultat dau	No n'hi ha	No en tenen
Cas 2	X = hores extra Y = sou	Y = sou fix + $A \cdot X$ A = € / hora extra	Relació molt forta Relació funcional
Cas 3	X = hores preparació Y = nota	No n'hi ha	Tenen relació, però no és tan forta com la funcional

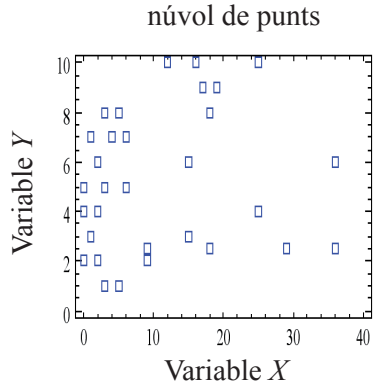
Com és evident, les relacions funcionals gaudeixen d'una fórmula que demostra el tipus de relació. Al contrari, per a la resta de parells de variables no hi ha cap fórmula absoluta, malgrat els lligams que existeixen en alguns casos. És per evidenciar-ho que sorgeix el concepte *correlació*, r_{XY}. Així, si dues variables tenen una relació lineal molt forta, el valor absolut de la correlació serà molt pròxim a 1. En cas contrari, serà pròxim a 0. Els casos 0 i 1 equivalen a no tenir cap tipus de relació i a tenir una relació funcional. El vector següent ho resumeix:

$|r_{xy}| = 0$ $\qquad$ $0 < |r_{xy}| < 1$ $\qquad$ $|r_{xy}| = 1$

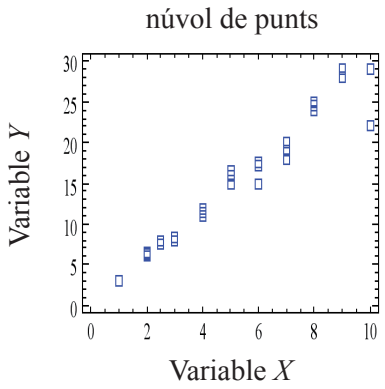
No existeix relació $\qquad$ Existeix relació $\qquad$ Relació funcional

Quan el tipus de relació funcional que s'estudia entre les variables és una funció lineal (una funció del tipus $y = ax + b$), es parla de *correlació lineal*. Al llarg de la unitat, quan es mencione el terme *correlació* es considerarà la correlació lineal, si no s'explicita una altra cosa.

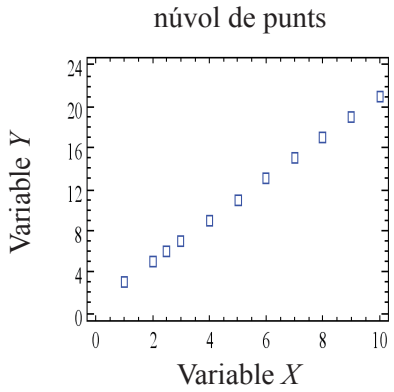
Una primera manera d'observar la relació existent entre les variables X i Y és mitjançant els gràfics de dispersió. Així, tenint en compte el que hem exposat al començament d'aquest punt sobre la correlació:



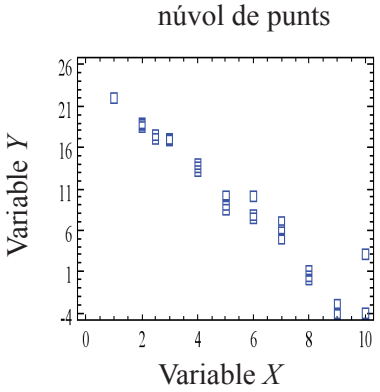
a) No existeix correlació



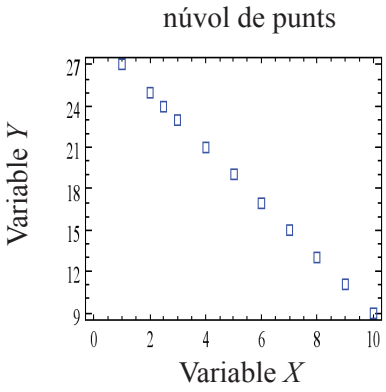
b) Correlació lineal positiva marcada



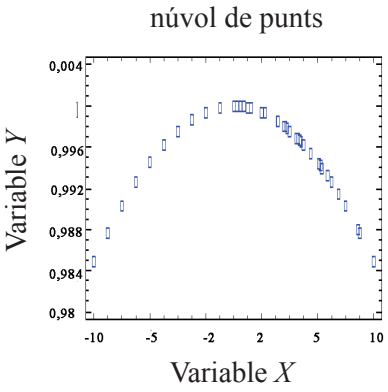
c) Correlació lineal positiva perfecta



d) Correlació lineal negativa marcada



e) Correlació lineal negativa perfecta



f) Correlació no lineal entre X i Y

Com es pot observar, en l'exemple *e* es detecta una relació entre les variables *X* i *Y*. No obstant això, és evident que no es tracta d'una relació lineal; ja que aquests tipus de relacions determinen un núvol de punts semblants a una línia recta (exemples *b*, *c*, *d* i *e*). En l'exemple *a*, no es distingeix cap tipus de lligam entre totes dues variables, els punts estan molt dispersos.

5.4.1. Covariància

El gràfic és una primera aproximació a l'estudi de la relació que existeix entre les variables, però únicament aporta informació de caire intuïtiu. El concepte que és necessari definir per a poder decidir si hi ha o no relació lineal entre dues variables és el de *correlació lineal*. En primer lloc, però, cal introduir el concepte de *covariància*.

La covariància és un estadístic (o un paràmetre) que es calcula semblantment al de variància i permet conèixer si dues variables estan relacionades o no linealment. Així, la fórmula que presentem té un sentit que passem a explicar:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(x_i - \overline{X})(y_j - \overline{Y}) \cdot n_{ij}}{n}.$$

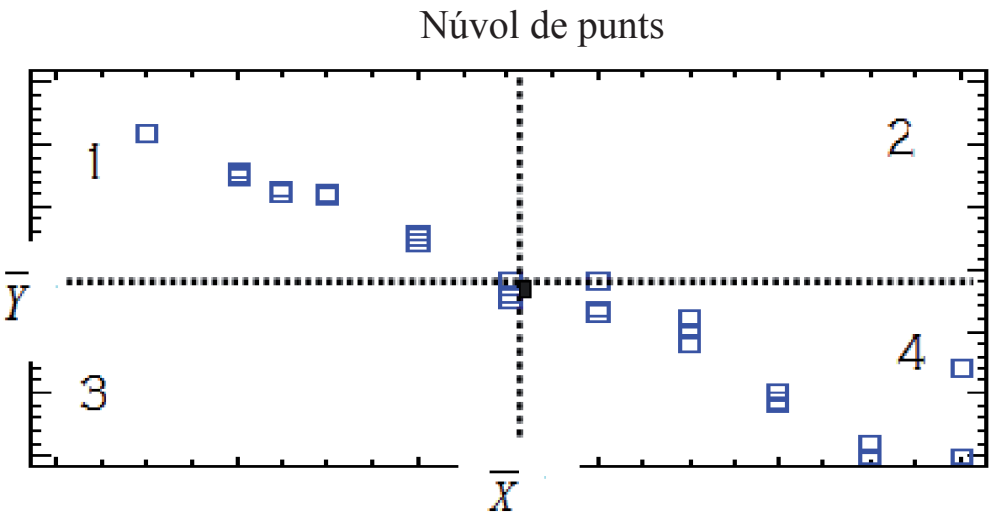


Figura 1

Si ara s'analitzen conjuntament la fórmula i el núvol de punts (figura 1), en el qual s'ha representat el punt $(\overline{X}, \overline{Y})$, es dedueix el següent:

- El punt $(\overline{X}, \overline{Y})$ divideix el núvol de punts en quatre quadrants en els quals s'ubiquen els punts.
- Per als punts que hi ha al quadrant 1 els valors $(x_i - \overline{X})$ són negatius, i els valors $(y_j - \overline{Y})$, positius. Justament els signes contraris tindran aquestes

diferències si els punts es troben al quadrant 4. Consegüentment, si majoritàriament es ditribueixen entre els quadrants 1 i 4 (gràfics *d* i *e*), els productes $(x_i - \overline{X}) \cdot (y_j - \overline{Y})$ són negatius i, òbviament, la suma de tots aquests productes també és negativa.

- Anàlogament, per als punts que hi ha al quadrant 2 els productes $(x_i - \overline{X}) \cdot (y_j - \overline{Y})$ són positius, ja que els factors ho són. El mateix ocorre si els punts estan al quadrant 3, ja que en aquest cas tots dos factors són negatius. Consegüentment, si majoritàriament els punts s'hi troben distribuïts entre els quadrants 2 i 3 (gràfics *b* i *c*), la suma de tots els productes $(x_i - \overline{X}) \cdot (y_j - \overline{Y})$ és positiva.
- Si els punts estigueren distribuïts entre els quatre quadrants, la suma dels productes $(x_i - \overline{X}) \cdot (y_j - \overline{Y})$ tindria valors positius i negatius que es podrien anul·lar entre si (gràfic *a*), i d'aquesta manera ser un nombre proper a 0.

Nota

Pot ocórrer que la covariància siga 0 i els punts seguisquen una determinada forma. Això és perquè la covariància estudia la relació lineal existent entre les variables, és a dir, pot ser que dues variables estiguen lligades entre si per una funció (gràfic *e*) i, en canvi, la covariància siga 0.

En resum:

- Si $S_{XY} > 0$ ==> dependència lineal directa o positiva, és a dir, quan augmenta la X augmenta la Y (exemples *b* i *c*).
- Si $S_{XY} = 0$ ==> incorrelades, és a dir, no hi ha relació lineal (exemples *a* i *f*).
- Si $S_{XY} < 0$ ==> dependència lineal inversa o negativa, és a dir, quan augmenta la X disminueix la Y (exemples *d* i *e*).

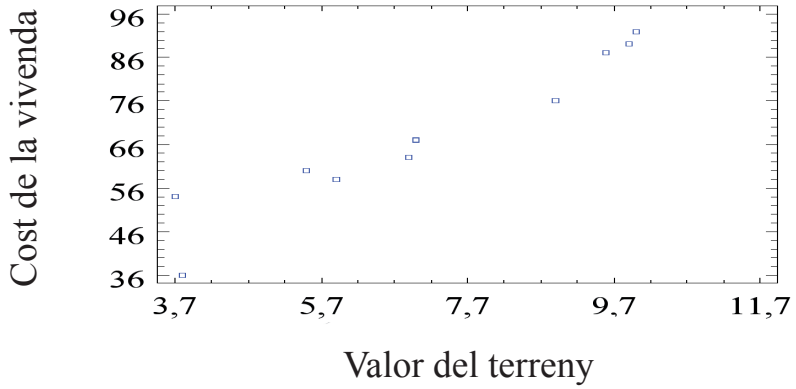
Exemple 6

Per al cas que s'està considerant en tota la unitat, exemple 1, calculem la covariància:

Sabent que *valor terreny* $= \overline{X} = 7{,}1586$; *valor habitatge* $= \overline{Y} = 68{,}0052$, i $n = 29$, es calcula:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(x_i - \overline{X})(y_j - \overline{Y})\cdot n_{ij}}{n} = \sum_{i=1}^{10}\sum_{j=1}^{11}\frac{(x_i - 7{,}1586)(y_j - 68{,}0052)\cdot n_{ij}}{29} = 42{,}1527\,€^2$$

Núvol de punts



Així doncs, el valor de la covariància és coherent amb el núvol de punts que s'ha obtingut: sembla que hi ha una relació directa o positiva entre el valor del terreny i el cost de l'habitatge. Com més valor del terreny, més cost de l'habitatge.

Propietats de la covariància:

Propietat 1
Si a tots els valors de la variable X, se'ls suma una constant b *i a tots els valors de la variable Y, una constant C, la covariància no varia.*

És a dir: $S_{X+b\,Y+C} = S_{XY}$

Se suposa que s'està treballant amb dues variables X i Y, de les quals es coneixen les mitjanes aritmètiques i la covariància.

Es defineix una altra variable X' a partir de la X sumant a cada valor de la variable X una constant b, és a dir, els valors de la variable X' són els generats per la variable X segons la relació $x'_i = x_i + b$, i una altra variable Y' a partir de la Y sumant a cada valor de la variable Y una constant C, és a dir, els valors de la variable Y' són els generats per la variable Y segons la relació següent: $y'_i = y_i + C$. En aquestes condicions és clar que es compleixen les propietats de la mitjana aritmètica. Llavors s'obté que $\overline{X'} = \overline{X} + b$ i $\overline{Y'} = \overline{Y} + C$.

Aplicant aquestes igualtats a la definició de covariància:

$$S_{X'Y'} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(x'_i - \overline{X'})(y'_j - \overline{Y'})\cdot n_{ij}}{n} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(x_i + b - (\overline{X} + b))(y_j + C - (\overline{Y} + C))\cdot n_{ij}}{n}$$

$$= \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(x_i - \overline{X})(y_j - \overline{Y})\cdot n_{ij}}{n} = S_{XY}$$

Exemple 7

La taula següent mostra els sous mensuals dels treballadors d'una empresa segons els anys que fa que hi treballen. Calcula si existeix o no dependència lineal entre ambdues variables l'any següent, si se sap que el sou augmentarà en 10 € per a cada treballador. D'altra banda, és sabut que aquest any la covariància ha estat de 3.260,299 anys · €.

Anys d'experiència	Sou	Treballadors
3	1.200	5
3	1.350	2
5	2.100	3
10	3.800	4

L'any següent, el nombre d'anys augmentarà en 1 per a cada treballador i el sou en 10€. Llavors, la taula de l'any següent queda:

Anys d'experiència	Sou	Treballadors
3 + 1 = 4	1.200 + 10 = 1.210	5
3 + 1 = 4	1.350 + 10 = 1.360	2
5 + 1 = 6	2.100 + 10 = 2.110	3
10 + 1 = 11	3.800 + 10 = 3.810	4

I en calular la covariància a partir d'aquesta taula es pot comprovar que és de 3.260,299 anys · €.

És a dir, no canvia. Aquest fet corrobora la propietat, ja que si prenem: X = anys experiència aquest any, X' = anys d'experiència l'any vinent, Y = sou aquest any i Y'= sou l'any vinent, es té que $X' = X + 1$ i $Y' = Y + 10$. Aplicant-hi la propietat, $S_{X+1\ Y+10} = S_{XY}$. Per tant, sí que hi ha una relació lineal positiva entre les variables.

Propietat 2
Si tots els valors d'una variable X *es multipliquen per una constant* a *i tots els valors de la variable* Y *per una constant* b*, la covariància queda multiplicada pel producte de les constants.*

És a dir: $S_{a\cdot Xb\cdot Y} = S_{XY}$.

Se suposa que s'està treballant amb dues variables *X* i *Y*, de les quals es coneixen les mitjanes aritmètiques i la covariància.

Es defineix una altra variable *X*' a partir de *X* multiplicant cada valor de la variable *X* per una constant *a* (és a dir, els valors de la variable *X*' són els generats per la variable *X* segons la relació següent: $x'_i = a \cdot x_i$), i una altra variable *Y*' a partir de *Y* multiplicant cada valor de la variable *Y* per una constant *b* (és a dir, els valors de la variable *Y*' són els generats per la variable *Y* segons la relació següent: $y'_i = b \cdot y_i$). En aquestes condicions és clar que es compleixen les propietats de la mitjana aritmètica. Llavors s'obté que $\overline{X}' = a \cdot \overline{X}$ i $\overline{Y}' = b \cdot \overline{Y}$.

Aplicant aquestes igualtats a la definició de covariància:

$$S_{X'Y'} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(x'_i - \overline{X}')(y'_j - \overline{Y}')\cdot n_{ij}}{n} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(ax_i - \overline{aX})(by_j - \overline{bY})\cdot n_{ij}}{n}$$

$$= \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{a\cdot(x_i - \overline{X})b\cdot(y_j - \overline{Y})\cdot n_{ij}}{n} = ab\sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(x_i - \overline{X})(y_j - \overline{Y})\cdot n_{ij}}{n} = abS_{XY}.$$

Exemple 8

Emprant les dades de l'exemple 7, calcula la covariància, per al cas que l'empresa decidira aquest mateix any augmentar un 10% els sous dels treballadors:

Si es pren *X* = anys experiència enguany, *X*' = anys experiència l'any vinent, *Y* = sou aquest any i *Y*'= sou l'any vinent, es té que $X' = X + 1$ i $Y' = 1{,}10 \cdot Y$.

Aplicant-hi la propietat $S_{X'Y'} = 1{,}10 \cdot S_{XY} \rightarrow S_{X'Y'} = 1{,}10\ 3.260{,}299 = 3.587{,}089 \cdot \text{anys} \cdot €$.

Propietat 3
A partir de les propietats anteriors: si es tenen dues variables X *i* Y *amb covariància* S_{XY}*, i dues transformacions lineals de les variables de la forma* X' = ax + c*, i* Y' = by + d*, la nova covariància es relaciona amb l'anterior de la forma:*

$$S_{X'Y'} = a \cdot b\ S_{XY}.$$

Se suposa que s'està treballant amb dues variables *X* i *Y*, de les quals es coneixen les mitjanes aritmètiques i la covariància.

Es defineix una altra variable X' a partir de X multiplicant cada valor de la variable X per una constant a i sumant-hi una constant c (és a dir, els valors de la variable X' són els generats per la variable X segons la relació següent: $x'_i = a \cdot x_i + c$), i una altra variable Y' a partir de Y multiplicant cada valor de la variable Y per una constant b i sumant-hi una constant d (és a dir, els valors de la variable Y' són els generats per la variable Y segons la relació $y'_i = b \cdot y_i + d$). En aquestes condicions és clar que es compleixen les propietats de la mitjana aritmètica. Llavors s'obté que $\overline{X'} = a \cdot \overline{X} + c$ i $\overline{Y'} = b \cdot \overline{Y} + d$.

Aplicant aquestes igualtats a la definició de covariància:

$$S_{X'Y'} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(x'_i - \overline{X'})(y'_j - \overline{Y'})\cdot n_{ij}}{n} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(ax_i + c - (a\overline{X} + c))(by_j + d - (b\overline{Y} + d)\cdot n_{ij}}{n} =$$

$$= \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{a\cdot(x_i - \overline{X})b\cdot(y_j - \overline{Y})\cdot n_{ij}}{n} = ab\sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(x_i - \overline{X})(y_j - \overline{Y})\cdot n_{ij}}{n} = abS_{XY}$$

Exemple 9

Emprant les dades de l'exemple 7, calcula la covariància per al cas que l'empresa decidira augmentar, l'any següent, en un 10% els sous dels treballadors:

Prenent X = anys experiència aquest any, X' = anys experiència l'any vinent, Y = sou aquest any i Y' = sou l'any vinent, es té que $X' = X + 1$ i $Y' = (1{,}10) \cdot Y$. Aplicant-hi la propietat $S_{X'Y'} = 1{,}10 \cdot S_{XY} \rightarrow S_{X'Y'} = 1{,}10\ 3.260{,}299 = 3.587{,}089 \cdot \text{anys} \cdot €$.

Exemple 10

Es consideren dues variables X i Y, de les quals es coneix que $S_{XY} = 3$. Llavors:

Propietat 1. Si es generen les variables $X' = X + 5$ i $Y' = Y + 6$, es té que $S_{X'Y'} = S_{XY} = 3$.

Propietat 2. Si es generen les variables $X'' = 2X$ i $Y'' = 5Y$, es té que $S_{X''Y''} = 2 \cdot 5 \cdot S_{XY} = 30$.

Propietat 3. Si es generen les variables $X''' = 2X + 5$ i $Y''' = 5Y + 6$, es té que $S_{X''Y''} = 2 \cdot 5 \cdot S_{XY} = 30$.

Nota

Existeix una altra forma d'obtenir la covariància, que és de càlcul més senzill:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{x_i y_j \cdot n_{ij}}{n} - \overline{X} \cdot \overline{Y}$$

Es pot demostrar l'equivalència d'ambdues definicions mitjançant procediments algebraics elementals. Es pot aconsellar, en termes generals, utilitzar la primera definició per a aspectes teòrics, com ara demostracions de propietats, i la segona per a càlculs en els exemples.

Exemple 11

Es veurà tot seguit un exemple d'aplicació d'aquesta darrera propietat a partir de la següent taula de doble entrada i es calcularà la covariància:

X \ Y	1,6	1,7	1,8	
60	2	1	0	3
70	2	4	2	8
80	1	1	4	6
90	0	2	1	3
	5	8	7	20

En primer lloc cal calcular la mitjana aritmètica de cada variable marginal:
$\overline{X} = 74{,}5$ i $\overline{Y} = 1{,}71$.

En segon lloc, el primer sumand de S_{XY}, $\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n}$. Per això, és necessari calcular primer els productes, sumar-los tots i després dividir el resultat pel nombre total de dades:

60 * 1,6 * 2 =	192
60 * 1,7 * 1 =	102
70 * 1,6 * 2 =	224
70 * 1,7 * 4 =	476
70 * 1,8 * 2 =	252
80 * 1,6 * 1 =	128
80 * 1,7 * 1 =	136
80 * 1,8 * 4 =	576
90 * 1,7 * 2 =	306
90 * 1,8 * 1 =	162
Total suma =	2.554

Per tant, es té que:

$$\sum_{i=1}^{h}\sum_{j=1}^{k} x_i \cdot y_j \cdot n_j = 2.554$$

$$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{2.554}{20}$$

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X} \cdot \overline{Y} =$$

$$= \frac{2554}{20} - 74{,}5 \cdot 1{,}71 =$$

$$127{,}7 - 127{,}395 = 0{,}305.$$

5.4.2. Correlació lineal

La covariància permet distingir si dues variables X i Y tenen una relació positiva, negativa o zero, però no aporta informació del grau de dependència d'una variable respecte de l'altra. A més a més, la covariància depèn de les unitats de mesura emprades per a X i Y, si per exemple X es mesura en m^3 i Y en mm^3, cada desviació de X augmenta S_{XY} 10^9 vegades. Per a fer front a aquesta i a altres qüestions es defineix el concepte de *correlació lineal* r_{XY}:

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y} \quad \text{on } S_X \text{ i } S_Y \text{ són les desviacions típiques de } X \text{ i } Y.$$

És evident que, per definició, el coeficient de correlació lineal informa de les mateixes coses que ho fa la covariància. A més a més, compleix una propietat molt important: està fitat per 1 i per –1. Així doncs, r_{XY} es caracteritza per:

- ser adimensional i estar comprès sempre entre –1 i 1,
- si hi ha relació lineal forta positiva, $r_{XY} > 0$ i és pròxim a 1,
- si hi ha relació lineal negativa forta, $r_{XY} < 0$ i és pròxim a –1,
- si no hi ha relació lineal r_{XY} serà 0.

Nota

- Amb posterioritat es justificarà més profundament el concepte i les propietats del coeficient de correlació.
- Quan les variables X i Y són independents, $S_{XY} = 0$, i per tant, $r_{XY} = 0$. És a dir, si dues variables són independents, la covariància val zero. No es pot assegurar el mateix en sentit contrari. Si dues variables tenen covariància zero, no es pot dir que són independents. Es pot dir que linealment no tenen relació, però podrien tenir un altre tipus de relació no lineal i no ser independents (exemple *f*).

Exemple 12

En l'exemple 1 que s'està considerant, per a calcular la correlació és necessari, de primer, conèixer les variàncies i la covariància. Aprofitant els càlculs anteriors, es té que: $S_{XY} = 42{,}1527$; $S_X = 18{,}1656$; $S_Y = 2{,}4242$. En conseqüència, el coeficient de correlació lineal $r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{42{,}1527}{18{,}1656 \cdot 2{,}4242} = 0{,}9572$. Per tant, la relació lineal entre totes dues variables és alta.

5.5. Recta de regressió. Bondat d'ajustament

Com s'ha exposat anteriorment, quan s'estudien dues característiques simultàniament sobre una mostra, es pot considerar que una influeix sobre l'altra d'alguna manera. L'objectiu principal de la regressió és descobrir la manera com es relacionen.

Per exemple, en una taula de pesos i altures de 10 persones (taula 3) es pot suposar que la variable «alçada» influeix sobre la variable «pes» en el sentit que els pesos grans estan explicats per valors grans d'alçada (en general).

Alçada	175	180	162	157	180	173	171	168	165	165
Pes	80	82	57	63	78	65	66	67	62	58

Taula 3

De les dues variables que s'han d'estudiar, que es denotaran amb X i Y, s'anomena *variable independent* o *explicativa* la X, i l'altra, Y, s'anomena *variable dependent* o *explicada*.

En la majoria dels casos la relació entre les variables és mútua, i és difícil saber quina variable influeix sobre l'altra. En l'exemple anterior, a una persona que mesura menys, se li suposa menor alçada i a una persona de poca alçada se li suposa un pes més petit. És a dir, es pot admetre que cada variable influeix sobre l'altra de forma natural i per igual. Un exemple més clar on distingir entre variable explicativa i explicada és aquell on s'anota, de cada alumne d'una classe, el seu temps d'estudi (en hores) i la seua nota d'examen. En aquest cas, un curt temps d'estudi tendirà a obtenir una nota més baixa, i una bona nota ens indicarà que, tal vegada, l'alumne ha estudiat molt. No obstant això, a l'hora de determinar quina variable explica l'altra, és clar que el temps d'estudi explica la nota d'examen i no al contrari, perquè l'alumne primerament estudia un temps, que pot decidir lliurement, i després obté una nota que ja no decideix a plaer. Per tant,

X = temps d'estudi (variable explicativa o independent)

Y = nota d'examen (variable explicada o dependent).

Malgrat tot, en general caldrà estudiar la situació que es planteja i, basant-se en aquesta anàlisi, decidir quina variable és la independent i quina la dependent.

D'altra banda, el problema de trobar una relació funcional entre dues variables és molt complex, ja que hi ha una infinitat de funcions diferents. El cas més senzill de relació entre dues variables és la relació lineal, és a dir, aquella en què $Y = a + b\,X$ (equació d'una recta), on a i b són qualsevol nombre real. Aquest és el cas que s'estudiarà.

Cal remarcar que el propòsit de tot el que s'ha esmentat en aquest punt fins ara no és altre que el d'introduir els conceptes més elementals de les funcions reals de variable real. L'objectiu principal de l'apartat és trobar la funció lineal que millor modele una distribució bidimensional de dades; és a dir, donat un conjunt de dades (x_i, y_i) s'estudia si tenen relació lineal entre si. Si és així es buscarà l'equació de la funció lineal que millor s'ajuste a les dades. D'aquesta manera es podran, fins i tot, fer previsions del valor corresponent a la variable dependent, per a un valor concret de la variable independent. La pregunta en aquest moment és com fer-ho.

Recta de regressió

Un dibuix del núvol de punts o diagrama de dispersió de la distribució pot indicar si és raonable pensar que pot haver-hi una bona correlació lineal entre totes dues variables (figura 2).

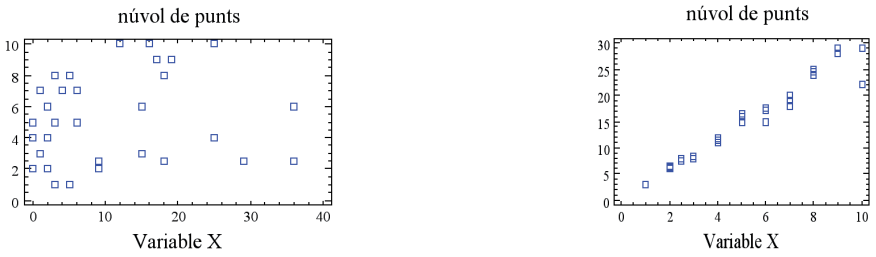


Figura 2

En els diagrames anteriors (figura 2) es pot observar que, en el de la dreta, una línia recta inclinada pot aproximar-se a quasi tots els punts, mentre que en l'altre, qualsevol recta deixa molts punts allunyats d'ella. Així, l'anàlisi de regressió lineal només estaria justificada en l'exemple de la dreta.

Com es pot veure en ambdós diagrames, cap recta és capaç de passar per tots els punts. De totes les rectes possibles, la recta de regressió de Y sobre X és aquella que minimitza l'error d'aproximació, considerant X com a variable explicativa o independent, i Y com l'explicada o dependent. Però, com es calcula la recta i es minimitza l'error?

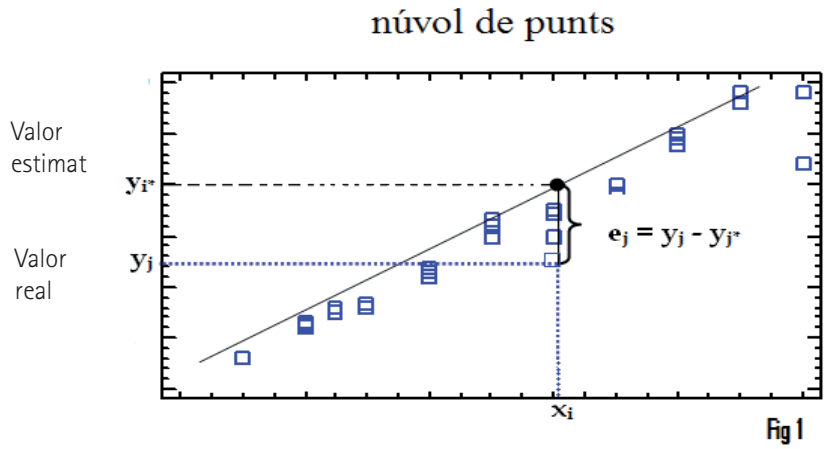
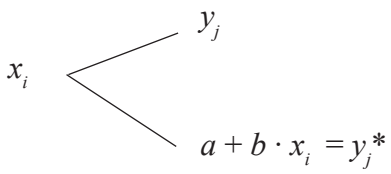


Figura 3

Es considera la recta $y = a + b\,x$, on a i b són paràmetres. D'aquesta manera, la recta o funció lineal és genèrica (representa totes les funcions lineals possibles, únicament cal donar valors als paràmetres per a obtenir les infinites rectes). El que es farà és trobar els valors dels paràmetres a i b de manera que la recta s'ajuste tant com siga possible als punts (figura 3). El mètode que s'empra per a buscar els valors dels paràmetres a i b és el dels mínims quadrats. Se'n farà un esbós seguidament.

Per a cada dada de X, és a dir, per a cada x_i del conjunt de dades (x_j, y_j), existeix una parella de Y, y_j (valor real), i també es pot calcular el valor que ix en substituir x_i en l'equació d'una recta genèrica, que es denotarà per y_{j*} (valor estimat):



Així, quan es pren la dada x_i l'error que es comet en triar y_{i*} en lloc del vertader y_i és la diferència, és a dir: $e_j = y_j - y_j^*$ (figura 3).

Aquests errors poden ser positius o negatius, i el que es fa és triar, de totes les rectes possibles, la recta que minimitze la suma dels quadrats de tots aquests errors, que és la mateixa que aquella que minimitza la variància dels errors.

És a dir, el mètode dels mínims quadrats busca els valors de a i de b que facen mínima la suma dels quadrats dels errors:

$$\sum_{i=1}^{h}\sum_{j=1}^{k} e_j^2 n_{ij} = \sum_{i=1}^{h}\sum_{j=1}^{k} (y_j - y_j^*)^2 n_{ij} = \sum_{i=1}^{h}\sum_{j=1}^{k} (y_j - a - bx_i)^2 n_{ij}.$$

Usant tècniques de derivació es dedueix que, de tots els possibles valors de a i de b, aquells que minimitzen la suma anterior són:

$$a = \bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x} \qquad \text{i} \qquad b = \frac{s_{xy}}{s_x^2}.$$

(Nota: cal no oblidar que si es coneixen les dades també es coneixen els termes: $\{\overline{Y}, \overline{X}, S_{xy}, S_x^2\}$ i per tant a i b seran nombres reals en el moment que es produïsquen les substitucions.)

Així, substituint en $Y = a + b\,X$, l'equació de la recta de regressió de Y sobre X és:

$$y = \left(\bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x}\right) + \left(\frac{s_{xy}}{s_x^2}\right) \cdot x$$

I recol·locant els termes es pot escriure de la forma següent, que correspon a l'equació d'una recta si en coneixem un punt i el pendent:

$$y - \overline{Y} = \frac{S_{XY}}{S_X^2} \cdot (x - \overline{X})$$

Si s'haguera pres *Y* com a variable independent o explicativa, i *X* com a dependent o explicada, la recta de regressió que es necessitaria seria la que minimitzara errors de la *X*. S'anomena *recta de regressió de X sobre Y* i es calcula fàcilment permutant els llocs de *x* i *y*, i s'obté:

$$x - \overline{x} = \frac{s_{xy}}{s_y^2} \cdot (y - \overline{y}).$$

Cal remarcar que totes dues rectes passen pel punt $(\overline{x}, \overline{y})$. Així, en la regressió de *Y* sobre *X*, quan *X* pren el valor $\overline{x}$ l'estimació de *Y* és $\overline{y}$.

El pendent de la recta de regressió de *Y* sobre *X* és $\frac{S_{XY}}{S_X^2}$ i el de *X* sobre *Y* és $\frac{s_{xy}}{s_y^2}$.

Atès que les variàncies són positives per definició, el signe dels pendents serà el mateix que el de la covariància i, així, les rectes seran ambdues creixents o decreixents, depenent de si la covariància és positiva o negativa, respectivament.

Nota

Cal advertir que la recta de regressió de *X* sobre *Y* no es calcula a partir de la recta de regressió de *Y* sobre *X* i després aïllant la *X*.

Exemple 13

A l'exemple 1 que s'està considerant, es té que la variable independent és el valor del terreny, i el valor de l'habitatge, la variable dependent.

Pels estudis realitzats al llarg de la unitat se sap que la relació és directa, ja que la covariància és positiva. Com que la correlació obtinguda ha estat un nombre proper a 1, la relació lineal entre les dues variables és marcada. Per tant, el càlcul de la recta de regressió té sentit. Per a calcular-la, s'utilitzarà la darrera expressió: $y - \overline{Y} = \frac{S_{XY}}{S_X^2} \cdot (x - \overline{X})$. Així, $\frac{S_{XY}}{S_X^2} = \frac{42{,}1527}{5{,}8768}$ i la recta serà $y - 7{,}1586 = \frac{42{,}1527}{5{,}8768} (x - 68{,}0052)$, i aïllant la variable *y*, l'equació de la recta queda: $y = 16{,}6583 + 7{,}17274\, x$.

Utilitzant el nom de les variables, la recta de regressió quedaria:

$$\text{Cost de l'habitatge} = 16{,}6583 + 7{,}17274 \cdot \text{valor del terreny}.$$

Bondat d'ajustament. Coeficient de determinació

Un núvol de punts que s'agrupa entorn d'una recta imaginària justifica l'estudi de la regressió lineal entre les variables, tanmateix la variable explicativa (d'ara endavant l'anomenarem *X*) no explica (valga la redundància) al 100% els resultats que s'observen en la variable explicada (d'ara endavant l'anomenarem *Y*).

L'únic cas en què una variable explica al 100% l'altra variable és aquell on els punts del núvol estan alineats. En aquest cas, el valor que li correspon per la recta a cada valor de la variable *X* coincideix amb el valor real. En general no és així i, per tant, convé tenir un control sobre la bondat d'ajustament de la recta teòrica de *Y* sobre *X* (o a l'inrevés). Aquest control està determinat per l'anomenat *coeficient de determinació lineal*.

Per altra part, si la variable *X* influeix en la variable *Y*, part de la variabilitat de *Y* quedarà explicada per la variabilitat de *X*, i una altra part per causes relacionades únicament amb *Y* o amb l'atzar. En aquest sentit convé tenir una mesura que indique el percentatge de variació de la variable *Y* que és explicat per la variable *X*. Recordant que l'estadístic que mesurava la dispersió no és altre que la variància, tot el que hem esmentat es pot resumir dient que la variància de *Y* està generada, d'una banda, per les dades de *X* (és a dir, per la variància de *X*), i d'una altra, per causes relatives a *Y* i a l'atzar.

El coeficient de determinació lineal es pot definir com el percentatge de variància de *Y*, que es pot explicar per *X*, i se sol anomenar *qualitat d'ajustament* o *bondat d'ajustament* perquè valora la proximitat del núvol de punts a la recta de regressió (o dit d'una altra manera, com està d'ajustat el núvol de punts a la recta de regressió).

Pel que fa al càlcul del coeficient de determinació, cal definir prèviament:

- La variància de la variable *Y*, que és explicada per la regressió lineal, anomenada S_r^2, i que representa la variabilitat de la variable *Y* causada per les variacions de la variable *X*.
- La variància residual, que es representa per S_e^2, determina en quina mesura difereixen els valors ajustats per la recta dels valors observats. És a dir, es planteja mesurar la magnitud dels residus.

Així:

$$S_r^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}(y_j^* - \overline{y^*})^2 \frac{n_{ij}}{n} \quad \text{i} \quad S_e^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}(y_j - y_j^*)^2 \frac{n_{ij}}{n}$$

Es pot demostrar matemàticament que en la regressió lineal de la variable *Y* sobre la variable *X*, la variància de la variable *Y* es pot descompondre de la manera següent:

$$S_Y^2 = S_r^2 + S_e^2$$

Així doncs, de la relació es dedueix que com més alta siga la variància explicada per la regressió lineal –S_r^2– respecte de la variància total, més petita serà la variabilitat de l'error d'ajustament –S_e^2– i millor serà la bondat d'ajustament.

Si ara es divideix l'expressió anterior per S_y^2, s'obté:

$$1 = \frac{S_r^2}{S_Y^2} + \frac{S_e^2}{S_Y^2}.$$

I reprenent el significat de *coeficient de correlació lineal* (R^2) com el percentatge de variància de *Y* que es pot explicar per *X*, es té:

$$R^2 = \frac{S_r^2}{S_Y^2} = 1 - \frac{S_e^2}{S_Y^2} \text{ (en tant per un).}$$

D'aquesta definició es poden traure algunes conclusions:

- $0 \leq R^2 \leq 1$, per ser la part d'un total.
- $R^2 = 1$ implica que la variància residual és nul·la i, per tant, l'ajustament és perfecte. En conseqüència, la relació entre totes dues variables és lineal.
- $R^2 = 0$ implica que la variància residual és igual a la variància de la variable Y i que la variable explicativa no aporta informació vàlida per a l'estimació de la variable explicada. En conseqüència, no existeix relació lineal entre les dues variables.
- Com més pròxim a 1 estiga R^2, millor serà la bondat o qualitat d'ajustament.

Per altra part, ja s'havia comentat en un epígraf anterior que el coeficient de correlació lineal aporta informació sobre el grau de la relació lineal entre les variables. Per tant, sembla clar que ha d'estar relacionat amb el de deteminació. Es veurà que aquest fet és del tot cert.

Se sap que:

$Y^* = a + b\,X$, on $b = \frac{S_{XY}}{S_X^2}$ (obtingut de la construcció de la recta de regressió).

La variància dels valors ajustats és igual, aplicant-hi les propietats de la variància:

$S_r^2 = b^2 S_X^2$ = (substituint b) = $\frac{S_{XY}^2}{\left(S_x^2\right)^2} S_x^2 = \frac{S_{XY}^2}{S_x^2}$, per tant el coeficient de correlació lineal pot expressar-se com:

$$R^2 = \frac{S_R^2}{S_y^2} = \frac{S_{XY}^2}{S_x^2 S_y^2}.$$

que, evidentment, coincideix amb el quadrat del coeficient de correlació lineal i justifica totes les propietats d'ambdós coeficients esmentades anteriorment:

$$r_{XY}{}^2 = R^2.$$

Es pot observar que una correlació forta de signe positiu o negatiu (per exemple, $r_{XY} = \pm 0{,}9$) dóna com a resultat el mateix $R^2 = 0{,}81$, i indica també una bona bondat d'ajustament.

Exemple 14

Si $R^2 = 86\%$ per a unes variables X i Y, es pot dir que la bondat d'ajustament és prou alta, encara que no se sap si la recta de regressió és creixent o decreixent.

Si es coneix el coeficient de correlació lineal, $r_{XY} = -0{,}77$, entre dues variables X i Y, se sap que la recta de regressió és decreixent (pel signe negatiu de r) i calculant $R^2 = r_{XY}{}^2 \cdot 100 = 59{,}29\%$ es té una bondat d'ajustament mitjana (no és molt pobra, però tampoc es pot qualificar de *bona*).

Prediccions. Usos i abusos

El primer objectiu de la regressió lineal era posar de manifest la relació existent entre dues variables estadístiques. Una vegada es constata que aquesta hi és i es calcula la recta de regressió apropiada, aquesta es pot usar per a obtenir valors de la variable explicada, a partir de valors de la variable explicativa.

Per exemple, si es comprova una bona correlació lineal entre les variables X = hores d'estudi setmanal i Y = nota de l'examen, amb una recta de regressió de Y sobre X que és:

$$Y = 0{,}9 + 0{,}6\, X.$$

es pot plantejar la pregunta: quina nota pot obtenir (segons les dades) un alumne que estudia 10 hores setmanals?

I la resposta és tan senzilla com calcular Y, substituint a l'equació de la recta $X = 10$, d'on resulta $Y = 6{,}9$. El coeficient de determinació és la dada que indicarà si la predicció obtinguda és fiable o no, ja que és el coeficient el que informa sobre la bondat d'ajustament.

En el moment de fer prediccions cal tenir certes precaucions, perquè és possible obtenir resultats absurds. Segons la recta de regressió anterior, un alumne que estudie 20 hores per setmana ($X = 20$) tindria un resultat de 12,9 punts en l'examen, la qual cosa no té sentit si s'avalua sobre 10. La limitació de la predicció consisteix en el fet que només es pot realitzar per a valors de X que estiguen situats entre els valors de X de la taula de dades inicials.

Exemple 15

En el cas que s'està considerant (exemple 1) des del principi de la unitat, es calcularà el coeficient de determinació per a conèixer la bondat d'ajustament, i es realitzarà una predicció.

El coeficient de determinació $R^2 = r_{XY}{}^2 = 0{,}9572^2 = 91{,}6231\%$. Aquest resultat indica que la recta s'ajusta prou bé a les dades.

D'altra banda, la recta de regressió era:

$$\text{Cost de l'habitatge} = 16{,}6583 + 7{,}17274 \cdot \text{valor del terreny}.$$

Per tant, si es volguera calcular el cost d'un habitatge sabent que el terreny val 7.200 euros, 7,2 en milers d'euros, aleshores el cost estimat de l'habitatge per la recta de regressió seria:

$$\text{Cost de l'habitatge} = 16{,}6583 + 7{,}17274 \cdot 7{,}2 = 68{,}3020 \text{ milers d'euros}.$$

Cal observar que el resultat és coherent amb les dades.

Exemple 16

Es realitzarà un estudi complet de l'exemple que es descriu al començament de l'epígraf. Es consideren les dades:

Alçada	175	180	162	157	180	173	171	168	165	165
Pes	80	82	57	63	78	65	66	67	62	58

Encara que en aquest cas es tenen dues variables molt relacionades, i no està clarament definit quina influeix sobre l'altra, es decideix estudiar com l'alçada dels individus influeix sobre el seu pes corporal. Llavors es pren X = alçada, com a variable explicativa i Y = pes, com a variable explicada.

En primer lloc cal representar el núvol de punts (figura 4), perquè informa si té sentit iniciar l'estudi de la regressió lineal o no.

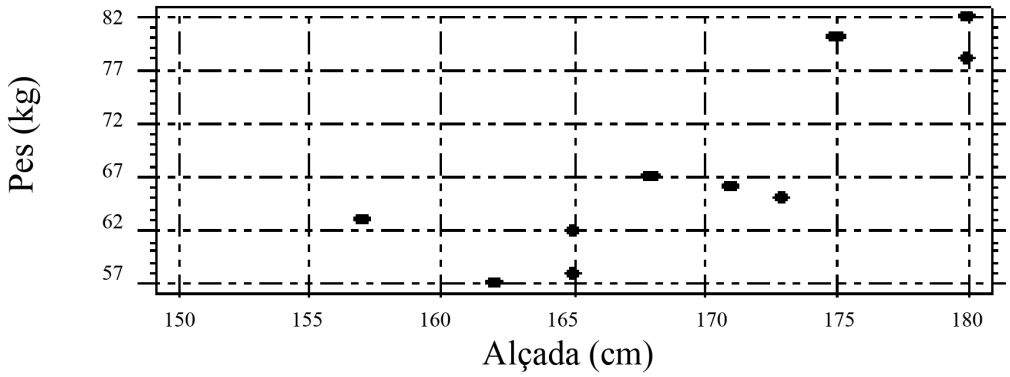


Figura 4

S'observa que els punts segueixen una tendència, encara que un, el 157,63, s'allunya de la dita tendència. Aquesta dada s'anomena *dada atípica*. En mostres nombroses una dada atípica no afecta massa el resultat i, de fet, algunes vegades s'elimina de la taula, encara que no es farà en aquest cas. Així, el dibuix revela certa tendència dels punts a agrupar-se entorn d'una recta imaginària. El coeficient de correlació, que és l'índex numèric que avaluarà aquesta tendència, constatarà que hi ha una bona relació lineal.

Es calculen els estadístics necessaris:

$$\overline{X} = 169{,}6 \qquad S_X = 7{,}2139$$
$$\overline{Y} = 67{,}8 \qquad S_Y = 8{,}7567$$

$$s_{xy} = \frac{175 \cdot 80 + 180 \cdot 82 + 162 \cdot 57 + \cdots}{10} - 169{,}6 \cdot 67{,}8 = 52{,}32.$$

Ara es pot calcular el coeficient de correlació lineal r_{XY} i el de determinació lineal R^2

$$R_{XY} = \frac{52{,}32}{7{,}2139 \cdot 8{,}7567} = 0{,}8282 \quad \text{i} \quad R^2 = (0{,}8282)^2 \cdot 100 = 68{,}59\%,$$

que indica que la variable independent «alçada» explica el 68,59% de la variància dels pesos. Aquest mateix coeficient de determinació es pren com a índex de fiabilitat a l'hora de fer prediccions de la variable «pes» a partir de dades de la variable «alçada».

Per exemple, segons la taula de dades, quin pes corporal hauria de correspondre a una persona de 178 cm d'estatura? La resposta ve de la recta de regressió de «pes» sobre «alçada». Es calcula amb les dades que es tenen,

$$y - 67{,}8 = \frac{52{,}32}{52{,}04} \cdot (x - 169{,}6)$$

i queda

$$y = -102{,}71 + 1{,}005x$$

Així, una persona d'alçada 178 cm (corresponent, per tant, a $X = 178$) té, en virtut de la recta de regressió, un pes (Y) que s'obté substituint el valor de X, i val $Y = 76{,}177$ kg. Es pren com a fiabilitat de la predicció l'índex R^2 calculat amb anterioritat. És a dir, es diu que la predicció té una fiabilitat del 68,59%.

Nota

Ja vam veure en l'exemple *f* que sovint la relació que hi ha entre les variables X i Y, una vegada vist el núvol de punts, suggereix un altre model de funció no lineal. Aquests casos s'aborden amb una metodologia semblant, que permet obtenir l'equació de la funció que millor s'ajusta a les dades, com per exemple funcions polinòmiques, racionals, exponencials, etc.

Aquests procediments queden fora del nostre propòsit en aquest text. Ara bé, el software disponible en aquests moments permet seleccionar el model que més convé, ja que ofereix, per a fer aquesta tria el coeficient R^2, de cadascun dels models que s'hi proposen. Tan sols cal triar el més adient considerant les característiques de la funció escollida (creixement local, discontinuïtats, asímptotes...) per a no tenir sorpreses amb les previsions en substituir un valor concret de X en l'equació de la funció que hem escollit.

5.6. Problemes proposats

En aquest epígraf es plantejaran un conjunt de problemes per a la resolució dels quals és necessari conèixer la teoria desenvolupada al llarg de la unitat.

Exercici 1

Per a realitzar un estudi sobre la utilització d'una impressora en un determinat departament es van mesurar els minuts transcorreguts entre les successives utilitzacions en un dia (X) i el nombre de pàgines impreses (I), i es van obtenir els resultats següents:

X	9	9	4	6	8	9	7	6	9	9	9	8	8	9	8	9
Y	3	8	3	8	3	8	8	8	3	8	12	12	8	8	8	12
X	9	9	10	9	15	10	12	12	10	10	12	10	10	12	12	10
Y	12	20	8	20	8	8	20	8	8	12	8	20	20	3	3	20

a) Escriu la distribució de freqüències conjunta. Quin és el percentatge de vegades que transcorren més de 9 minuts des de l'anterior utilització i s'imprimeixen menys de dotze pàgines? Quantes vegades s'imprimeixen menys de dotze pàgines i transcorren 9 minuts des de l'anterior utilització?
b) Quantes vegades s'imprimeixen com a màxim dotze pàgines? Quantes pàgines s'imprimeixen com a màxim en el 80% de les ocasions?
c) Troba la distribució de freqüències del nombre de pàgines impreses condicionada al fet que han transcorregut 9 minuts entre successives utilitzacions.
d) Dibuixa el diagrama de dispersió.

Exercici 2

El propietari d'un local de música pensat per a persones d'uns quaranta o cinquanta anys vol realitzar un estudi sobre l'edat dels clients i el consum que realitzen al local. La taula següent en mostra els resultats.

Ed.	50	51	53	50	51	48	50	49	52	52	49	50	52	51	52	49
Co.	3,2	4,1	4,5	3	3,6	2,9	3,8	3,8	3,6	3,9	3	3,8	4,1	3,5	4,0	3,1
50	51	50	51	52	53	52	52	51	50	51	54	50	51	51	51	52
3,1	4,3	3,3	3,9	3,7	4,1	4,2	3,5	3,8	3,6	3,4	4,6	3,5	3,6	3,1	4	3,8
51	52	51	50	51	49	51	48	50	52	53	52	50	52	51	51	51
4,2	4	4,4	3,9	3,7	3,4	3,3	2,7	3,4	3,6	4,4	4,3	3,3	4,2	4,2	3,3	3,7

a) Construeix la taula de doble entrada agrupant el consum en intervals d'amplaràries 0,5 i de manera que l'extrem petit de la primera classe siga 2,5.
b) La mitjana d'edat de les persones que acudeixen al local és l'esperada pel propietari? És aquesta mitjana representativa de les dades?
c) Quin és el consum mitjà dels clients que tenen 51 anys? Quin percentatge d'aquests consumeix menys de 4 €?
d) Quina és la mitjana d'edat dels clients que consumeixen entre 3,5 i 4 euros? Quin percentatge d'aquests té 50 o 51 anys?
e) Quin percentatge de clients consumeix 4 o més euros i té més de 50 anys?
f) Quin percentatge de clients té més de 51 anys o consumeix 4 o més euros?
g) Calcula i interpreta la covariància. Estan correlades les dues variables?
h) Representa el núvol de punts de la distribució de dades conjunta.

Exercici 3

Es van recol·lectar els valors mensuals de les despeses en publicitat d'una companyia ferroviària i el nombre de passatgers per als 15 mesos més recents. Les dades, les mostra la taula següent:

Publicitat (en milers)	10	12	8	17	10	15	10	14	19	10	11	13	16	10	12
Passatgers (en milers)	15	17	13	23	16	21	14	20	24	17	16	18	23	15	16

a) Calcula la despesa mitjana i el nombre mitjà de passatgers.
b) Calcula la covariància i el coeficient de correlació lineal. Interpreta'ls.
c) Si per a l'any següent es preveu un augment del 10% en la despesa de publicitat i, a més a més, l'empresa ha de pagar en concepte de cànon 15.000 € a una altra empresa, quina serà la despesa mitjana en publicitat? D'altra banda, es preveu un augment del 15% de passatgers l'any següent. Quina serà la covariància aleshores? I el coeficient de correlació lineal?

Exercici 4

En una mostra de 1.500 empreses es recullen dades sobre el nombre de treballadors de l'empresa (X) i la facturació anual en milions d'euros (Y). Els resultats es mostren resumits en els estadístics següents:

$\overline{X}$ = 14 treballadors $\overline{Y}$ = 100 millions S_X = 2 treballadors S_Y = 25 milions
S_{XY} = 45 treballadors · milió

a) Calcula la correlació lineal i interpreta-la.
b) Calcula el model de regressió lineal que millor aproxima la facturació en funció del nombre de treballadors.

c) En funció d'aquest ajustament calcula de manera aproximada la quantitat que s'espera que facture una empresa amb 15 treballadors. És fiable aquesta predicció? Raona la resposta.
d) Calcula el model de regressió lineal que millor aproxima el nombre de treballadors en funció de la facturació.
e) En funció d'aquest ajustament calcula de manera aproximada el nombre de treballadors que s'espera que tinga una empresa que factura 105 milions. És fiable aquesta predicció? Raona la resposta.
f) L'any següent, totes les empreses de la mostra augmenten en 8 treballadors la plantilla. Com a conseqüència, totes les empreses augmenten en un 10% la facturació. Respon a les mateixes qüestions plantejades als apartats *a*, *b* i *c* anteriors amb aquestes condicions.

Exercici 5

Els inversors en accions tendeixen a diversificar les seues carteres de valors. Una entitat bancària desitja comprovar si aquesta tendència es dóna entre vuit dels seus clients. Per a comprovar-ho, es realitza una recerca sobre els diferents tipus d'accions i els milions d'euros que hi han invertit. S'obtenen els resultats següents:

Tipus d'acció		12	8	10	11	7	7	10	14
Milions d'euros	58	42	51	54	40	39	49	56	

a) Representa el núvol de punts.
b) Existeix una associació lineal entre les dues variables?
c) Calcula les dues rectes de regressió.
d) Troba la bondat d'ajustament.

Exercici 6

La despesa pública (en milers d'euros) en concerts i subvencions per administracions educatives i anys, en el període 1993-2002, està donada per la taula següent:

	1995	1996	1997	1998	1999	2000	2001	2002 (4)
TOTAL	1.992.297	2.070.223	2.207.328	2.385.174	2.678.917	2.923.379	3.207.373	3.505.166
Andalusia	287.741	302.207	310.681	326.603	360.160	377.443	409.595	434.544
Canàries	42.594	46.963	49.854	56.766	62.956	67.932	72.176	77.015
Catalunya	385.237	412.144	486.590	509.252	581.831	604.021	638.958	690.919

	1995	1996	1997	1998	1999	2000	2001	2002 (4)
Comunitat Valenciana	174.531	183.746	192.339	218.842	259.208	289.563	362.629	?
Galícia	95.384	98.947	101.576	105.876	123.388	138.906	152.837	161.244
Navarra (Comunitat foral de)	49.291	51.408	54.386	56.541	61.674	67.569	74.614	78.868
El País Basc	293.052	295.526	306.037	327.179	339.117	351.003	388.033	428.694

Si se sap que l'any 2002 la despesa a Catalunya fou de 690.919 milers d'euros, es pot conèixer una aproximació de la de despesa de la Comunitat Valenciana?

TEMA 6

Introducció a la probabilitat (I): conceptes elementals

Objectius tema 6

- Distingir entre fenòmens deterministes i aleatoris.
- Identificar els fenòmens governats per l'atzar i comprendre el concepte intuïtiu de *probabilitat*.
- Saber emprar la teoria de conjunts per a representar esdeveniments o successos i les seues relacions.
- Conèixer els principis axiomàtics de la teoria de la probabilitat.
- Calcular la probabilitat de diferents esdeveniments, així com aplicar el teorema de la probabilitat total i el teorema de Bayes a les situacions que ho requerisquen.
- Saber identificar esdeveniments independents i calcular probabilitats d'aquest tipus d'esdeveniments.

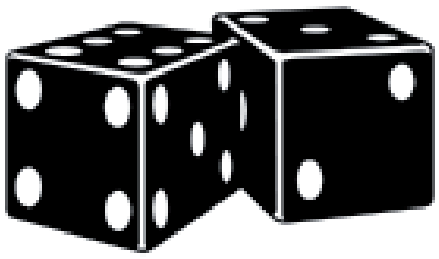
6.1. Introducció

La teoria de la probabilitat té l'origen en els jocs d'atzar, que es van convertir en una activitat molt comuna en la França del segle XVII. En aquest tipus de jocs (cartes, daus...) s'apostaven grans quantitats de diners, per això l'interès a predir-ne els resultats.

Un ric jugador professional parisenc de l'època, Antoine Gombaud (*Cavaller de Méré*), va plantejar un problema referent al joc d'atzar anomenat *points* a un dels matemàtics coetanis, Blaise Pascal. En aquest joc es compten els punts guanyadors en una tirada de daus, i qualsevol jugador que siga el primer a guanyar un cert nombre de punts és el vencedor i s'endú els diners de l'aposta.

Segons conten, Gombaud s'havia trobat jugant a *points* amb un jugador més experimentat quan, a causa d'un compromís, es veieren obligats a deixar el joc a mig fer. El problema que es plantejà aleshores fou què es podia fer amb els diners que s'havien apostat. La solució simple hauria estat donar tots els diners al competidor amb més punts, però Gombaud demanà a Pascal, si hi havia una manera més justa de dividir els diners. Demanaren a Pascal, doncs, que calculara la probabilitat que quedava a cada jugador de guanyar, si haguera continuat el joc i partint de la base que tots dos jugadors tenien les mateixes probabilitats de guanyar els punts que quedaven en joc. Els diners de l'aposta es dividirien segons aquest càlcul de probabilitats.

Així, qüestions com l'anterior o de l'estil: «Per què en el joc consistent en llançar un dau dues vegades i sumar les dues puntuacions és més senzill obtenir un vuit que un quatre?» o «Per què és més senzill obtenir una cara en llançar dues vegades una moneda que obtenir-ne dues en llançar la moneda quatre vegades, si pareix clar que ha de resultar igual de senzill raonant en termes de proporcionalitat?», centraren els primers estudis de probabilitat en aquesta època de la història.



Abans del segle XVII, les lleis de probabilitat eren definides per la intuïció i per l'experiència dels jugadors, però Pascal inicià un intercanvi epistolar amb Fermat, amb el fi d'esbrinar les regles matemàtiques que descriuen la probabilitat. Tres segles més tard, Bertrand Russell comentaria aquest aparent oxímoron: «Com es pot parlar de *lleis de la probabilitat*? No és la probabilitat l'antítesi de la llei?».

Els problemes de probabilitat de vegades són controvertits, perquè la resposta matemàtica, la vertadera resposta, sol ser contrària al que la intuïció sol suggerir. Per exemple, un dels problemes més contraintuïtius que hi ha es refereix a la probabilitat de celebrar l'aniversari el mateix dia que alguna altra persona. Si imaginem una festa amb 23 persones, no sembla molt probable que dues persones qualssevol complisquen anys el mateix dia. Amb 23 persones i 365 dies per a triar, la intuïció s'apropa a la idea que ningú comparteix la seua data d'aniversari. Si se'ns demanara posar una xifra a aquesta probabilitat, molts conjecturaríem una probabilitat de, potser, un 0,1. Doncs bé, la resposta real és de més del 0,5. Això vol dir que, posades en la balança de les probabilitats, és més probable trobar dues persones en la festa que compartisquen data d'aniversari que al contrari.

No és estrany, doncs, que els matemàtics de l'època es deixaren seduir per aquest tipus de problemes que, en moltes ocasions semblen anar en contra de la intuïció. De fet, Fermat i Pascal descobriren les regles essencials que governen tots els jocs de probabilitat, i que poden ser utilitzats pels jugadors per a definir el joc perfecte i les estratègies de les seues apostes. Més encara, aquestes lleis de probabilitat han trobat aplicacions en un gran nombre de situacions, que van des de les especulacions en el mercat de valors fins a les estimacions de la probabilitat d'un accident nuclear. Pascal (Singh, 1997) estava convençut, fins i tot, que podia utilitzar les seues teories per a justificar la creença en Déu. Va afirmar que «l'excitació que un jugador sent quan fa una aposta és igual a la quantitat que pot guanyar multiplicada per la probabilitat de guanyar-la». Aleshores va sostenir que el possible premi de la felicitat eterna té un valor infinit i que la probabilitat d'arribar al cel i tenir una vida virtuosa, per més petita que aquesta siga, és certament finita. Aleshores, segons la definició de Pascal, la religió era un joc d'excitació infinita i que valia la pena jugar-hi, perquè multiplicar un premi infinit per una probabilitat finita dóna un resultat infinit.

La correspondència epistolar que Pascal va mantenir amb Pierre Simon de Fermat i altres grans matemàtics de l'època originà la teoria de la probabilitat i féu que aquesta passara de ser una mera col·lecció de problemes aïllats sobre jocs, a constituir, amb el temps, una part molt important de les matemàtiques.

6.2. Atzar i probabilitat

En essència, el concepte inherent als jocs tractats en la introducció n'és bàsicament un, l'atzar:

Què s'entén per *atzar* i per *experiment aleatori*?

Per a entendre completament aquest concepte és convenient comparar-lo amb la idea oposada: la necessitat.

Així, els fenòmens governats per la necessitat són aquells en què una mateixa causa determina, inevitablement, un efecte. Per exemple:

- Una pedra cau si res la suporta.
- Si un líquid es calfa prou s'evapora.
- Si un nombre enter és divisible per 4 també ho és per 2.
- Si se sumen els angles d'un triangle el resultat és 180^{o}.

Aquestes afirmacions poden considerar-se com a casos de necessitat, ja que mai s'ha observat que una pedra sense suport no caiga, i és sabut que si un nombre és múltiple de 4 també ho és de 2. D'altra banda, els exemples anteriors mostren que la necessitat pot ser empírica (observacions) o lògica (basada en axiomes).

Com a conclusió, es pot asseverar que en els fenòmens dominats per la necessitat, quan ocorren certes causes l'efecte està determinat; per això, aquests fenòmens s'anomenen *deterministes*.

Enfront d'aquests fenòmens se'n perceben d'altres que atribuïm a l'atzar:

- Resultat obtingut en llançar una moneda sobre una taula.
- Resultat de llançar un dau.
- El nombre de telefonades que es reben en una centraleta.
- Resultat obtingut en un sorteig d'una panera de Nadal.
- ...

En aquest tipus d'experiments, si s'anomenen *esdeveniments* els possibles resultats del fenomen (per exemple, cara i creu són els esdeveniments de l'experiment «llançar una moneda»), un esdeveniment A_1 pot ocórrer quan es donen una sèrie de condicions, però sota aquestes mateixes condicions pot ocórrer qualsevol altre esdeveniment A_2, A_3, ... A_n dels possibles. És a dir, no pot preveure's quin de tots succeirà. En aquest cas es diu que l'ocurrència A_1 ha estat causada per l'atzar i que el fenomen és aleatori.

Per exemple, en l'experiment «llançar un dau» els possibles esdeveniments són: que isca un 1, que isca un 2, que isca un 3, que isca un 4, que isca un 5 i que isca un 6.

Si el resultat del llançament ha sigut 6, és evident que si es realitza l'experiment novament, el resultat podria ser 6, o qualsevol altre resultat dels possibles: 1, 2, 3, 4 o 5. Per això es diu que el resultat ha sigut conseqüència de l'atzar i l'experiment es denomina *aleatori*.

Què s'entén per *probabilitat*?

La necessitat està lligada a la certesa o seguretat, de fet s'afirma que:

- És segur que una pedra sense suport cau.
- És cert que la Terra es mou en l'espai.

De manera semblant l'atzar s'associa amb la noció de probabilitat, per això es diu que:

- És probable que puge la borsa.
- És probable que un equip de futbol guanye la lliga.

En aquest sentit, la probabilitat expressa intuïtivament una apreciació de la facilitat que s'atribueix que ocórrega un esdeveniment, o una mesura del grau de creença o versemblança en l'aparició d'un esdeveniment.

A més, hi ha una tendència més o menys conscient en l'ésser humà a pensar que un esdeveniment és més versemblant que un altre i, per tant, un desig de mesurar aquesta creença:

- Si s'extrau una bola d'una urna en què hi ha tres boles blanques i una de negra, és més probable que s'obtinga una bola blanca.
- És més probable que es consumisca més energia un divendres que un dissabte.
- ...

Com es veu, és relativament senzill comparar la versemblança de dos esdeveniments si es coneix la naturalesa de l'experiment aleatori.

D'altra banda, cal tenir present que l'objectiu fonamental és comparar tots els possibles esdeveniments o resultats d'un experiment segons la seua versemblança. El mètode més simple consisteix a assignar un nombre a cada un d'aquests possibles resultats, els quals han de permetre realitzar la dita comparació.

Es busca, doncs, una llei que governe l'atzar però que respecte el seu caràcter imprevisible.

Evidentment, d'ara endavant es considerarà que els experiments són aleatoris, ja que si foren deterministes no tindria sentit parlar de *probabilitat*.

6.2.1. Concepte de *probabilitat* basat en la freqüència

La manera més intuïtiva d'assignar un nombre a cada esdeveniment consisteix a realitzar l'experiment successives vegades i estudiar els resultats de les observacions. Així, per exemple, després de llançar una moneda 100 vegades s'han obtingut els resultats següents:

C, *C*, +, +, +, *C*, *C*, *C*, +, *C*, *C*, *C*, *C*, +, +, *C*, *C*, +, +, +, +, *C*, +, *C*, +, *C*, *C*, +, *C*, +, +, *C*, +, +, *C*, *C*, *C*, *C* +, +, +, +, +, *C*, *C*, +, +, *C*, +, +, *C*, +, *C*, +, +, +, *C*, *C*, *C*, *C*, *C*, +, *C*, *C*, *C*, +, +, +, *C*, *C*, *C*, +, *C*, *C*, *C C*, *C*, *C*, +, *C*, *C*, +, *C*, +, +, +, +, *C*, *C*, +, *C*, +, +, *C*, +, +, *C*, *C*, *C*, +, +

Els dos gràfics següents reflecteixen la diferència existent entre el nombre de cares i el de creus en cada llançament. S'hi observa l'aparició de ratxes de cares i creus, les quals provoquen que la dita diferència no s'estabilitze entorn del 0, com pareix indicar la intuïció.

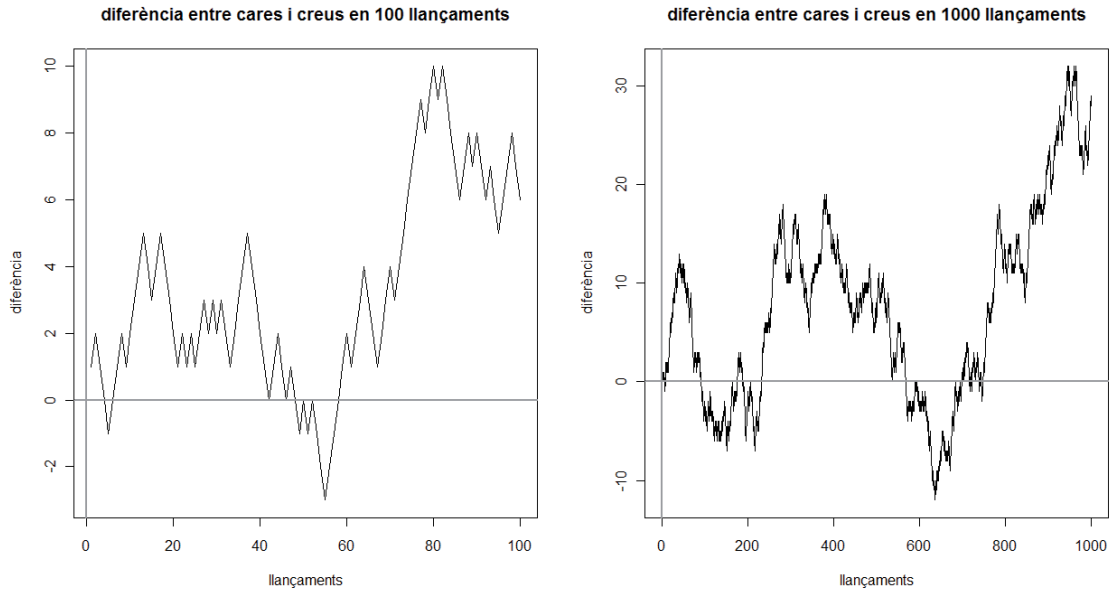


Figura 1

Vol dir açò que la nostra intuïció és errònia? És més probable que resulten més cares que creus en l'experiment? Si se centra l'atenció en la figura 2, reflex de la freqüència relativa del nombre de cares en cada llançament, s'observa que aproximadament a partir del llançament número 200 (en el gràfic dels 1.000 llançaments) la freqüència relativa del resultat «eixir cara» s'estabilitza entorn de 0,5. És més, pareix clar que a mesura que es va reiterant l'experiment amb nous llançaments, la freqüència relativa s'aproxima ràpidament a un valor ideal de 0,5.

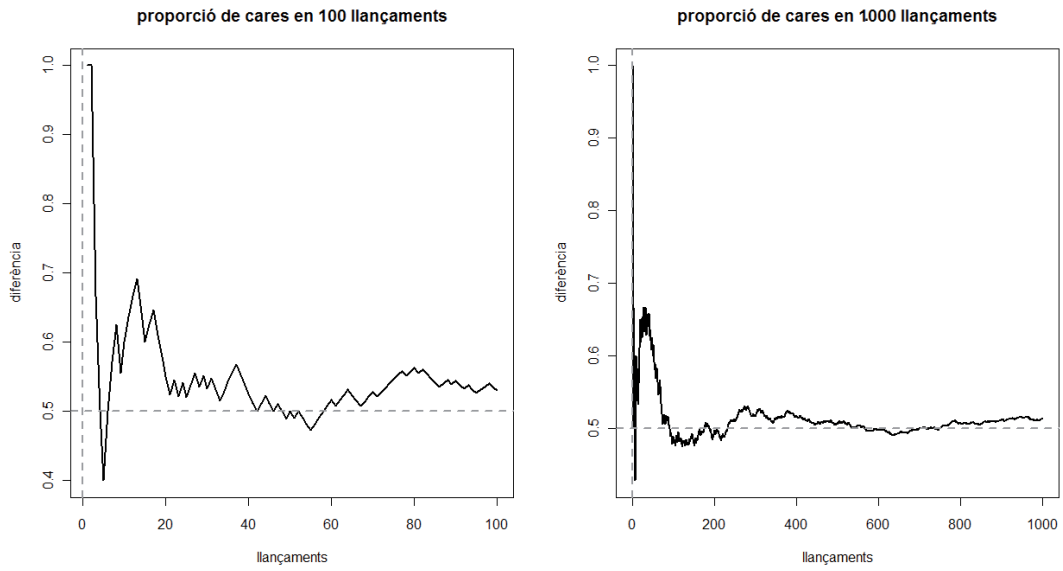


Figura 2

Cap a quin valor ideal s'aproximarà la freqüència relativa del resultat «eixir creu»? Òbviament, cap a 0,5.

D'altra banda, l'aparició de ratxes demostra que els resultats de l'experiment són totalment imprevisibles, encara que les freqüències relatives s'acosten a un valor ideal de 0,5.

Tenint en compte les diferents consideracions anteriors, s'obté una primera noció de probabilitat com a mesura de versemblança d'un esdeveniment. Així, es pot definir la probabilitat d'un esdeveniment $A1$ –d'ara endavant $P(A1)$– com el valor ideal a què s'acosta la freqüència relativa de l'esdeveniment A1 quan es realitza l'experiment un gran nombre de vegades.

Per ser la probabilitat de A1 una freqüència relativa, es té que: $0 \leq P(A1) \leq 1$.

Aquest fet queda ratificat per la llei dels grans nombres demostrada pel cèlebre matemàtic Jakob Bernoulli, que diu així: «La freqüència relativa d'un esdeveniment tendeix a estabilitzar-se entorn d'un nombre a mesura que el nombre de proves de l'experiment creix indefinidament».

En l'exemple P (eixir cara) = freqüència relativa (eixir cara) en moltes repeticions de l'experiment, i com s'ha vist al gràfic, la dita freqüència és 0,5.

Cal notar, que segons el tractament que s'ha donat al concepte, la probabilitat que isca cara o que isca creu, és una conseqüència lògica d'una llei empírica que podria enunciar-se de la manera següent: «Si es llança una moneda nombroses vegades, la freqüència relativa del resultat ''eixir cara'', s'aproxima a un valor denominat *P (eixir cara)* = 0,5».

Per tant, *P* (eixir cara) i *P* (eixir creu) passen a ser dues qualitats objectives més de la moneda, com la massa, el color..., ja que, de la mateixa manera que no es pot afirmar que la moneda no és sotmesa a l'acció de la gravetat en cap moment, tampoc es pot afirmar que si es repeteix l'experiment de llançar la moneda infinites vegades la freqüència de *P* (eixir cara) no s'acoste a 0,5.

6.2.2. Concepte subjectivista de *probabilitat*

En la interpretació de *probabilitat* com una freqüència és necessari repetir l'experiment un considerable nombre de vegades per a saber a quin número s'acosta la freqüència relativa de l'esdeveniment que s'està estudiant. No obstant això, en la majoria de les ocasions és convenient conèixer-ne la probabilitat abans que succeïsca. Per exemple, per no poder-se realitzar l'experiment (calcular la probabilitat que el nostre bitllet de loteria siga el premiat).

És per això que sorgeix la idea d'assignar un nombre a la probabilitat d'un esdeveniment segons la creença que es tinga que ocórrega. Cal notar que el nombre assignat ha d'estar comprès entre 0 i 1 perquè siga compatible amb la definició basada amb la freqüència de probabilitat comentada anteriorment.

Per exemple, davant de l'experiment «llançar una moneda», es pot assignar *P* (eixir cara) = 0,5 i *P* (eixir creu) = 0,5, ja que es té la creença que tots dos esdeveniments poden donar-se amb la mateixa facilitat.

És clar que les assignacions de probabilitats han de realitzar-se de manera que les dues interpretacions de la probabilitat coincidisquen; és a dir, l'assignació de la probabilitat realitzada a priori d'un esdeveniment ha de coincidir amb la probabilitat a posteriori. Una vegada assignades les probabilitats als esdeveniments –tant si és pel mètode empíric com si és pel subjectiu– les diferents deduccions que es realitzen segueixen un procediment lògic.

Assignació de probabilitats

Quan s'han d'assignar probabilitats als esdeveniments d'un experiment aleatori, és crucial entendre que el grau de versemblança depèn de l'evidència de la qual es disposa sobre el fenomen.

Exemple 1

Experiment: llançar un dau

Esdeveniments més simples: {eixir 1, eixir 2, eixir 3, eixir 4, eixir 5, eixir 6}.

Pareix clar que en aquest cas, tots els resultats poden donar-se amb la mateixa facilitat, per això:

$$P(\text{eixir } 1) = \frac{1}{6}\ P(\text{eixir } 2) = \frac{1}{6}\ P(\text{eixir } 3) = \frac{1}{6}\ P(\text{eixir } 4) = \frac{1}{6}\ P(\text{eixir } 5) = \frac{1}{6}$$
$$= P(\text{eixir } 6) = \frac{1}{6}$$

A més a més, es pot calcular la probabilitat d'experiments compostos:

$$P(\text{eixir parell}) = \frac{3}{6} = \frac{1}{2} \qquad P(\text{eixir més alt que } 4) = \frac{2}{6} = \frac{1}{3}.$$

Exemple 2

Experiment: llançar dues monedes equilibrades

Esdeveniments més simples: {eixir 2 cares, eixir 2 creus, eixir 1 cara i 1 creu}.

Una mala interpretació de l'experiment faria pensar que els tres resultats simples de l'experiment es donen amb la mateixa facilitat, i per tant s'hi assignarien, errò-niament, les probabilitats:

$$P(\text{eixir 2 cares}) = P(\text{eixir 2 creus}) = P(\text{eixir 1 cara i 1 creu}) = \frac{1}{3}.$$

Un estudi més reflexiu del fenomen que s'està estudiant ens porta a la conclusió que els resultats més simples són: {(cara, cara), (creu, creu), (cara, creu), (creu, cara)}, ja que el resultat «eixir cara en la primera moneda i creu en la segona» és diferent del resultat «eixir creu en la primera moneda i cara en la segona».

És per això que les probabilitats dels esdeveniments anteriors són:

$$P(\text{eixir 2 cares}) = P(\text{eixir 2 creus}) = \frac{1}{4} \qquad P(\text{eixir 1 cara i 1 creu}) = \frac{2}{4} = \frac{1}{2}.$$

Com es veu, és molt important la plena comprensió de l'experiment en el procés d'assignació de probabilitats.

6.3. Espai mostral i esdeveniments

Als epígrafs anteriors, davant d'un experiment s'han observat intuïtivament esdeveniments elementals i compostos. Formalitzarem aquests conceptes, així com les relacions que s'hi poden establir.

6.3.1. Definicions

Experiments aleatoris: es diu que un experiment és aleatori, estocàstic o estadístic quan, podent-se repetir indefinidament en condicions anàlogues, és impossible predir-ne el resultat, encara que es coneguen les condicions inicials.

En un experiment aleatori no es coneix el resultat fins que s'ha realitzat la prova. S'anomena *prova* cada realització d'un experiment.

Espai mostral: el conjunt de tots els resultats possibles que pot donar lloc un experiment aleatori s'anomena *espai mostral*. Sol representar-se per E o $\varnothing$ i es diu que és finit si el nombre de resultats possibles és finit.

Esdeveniment: donat un experiment aleatori, l'espai mostral del qual és E, s'anomena *esdeveniment* cada un dels subconjunts de E.

Es distingeixen els tipus d'esdeveniments següents:

- *Esdeveniment elemental*: només consta d'un element.
- *Esdeveniment compost*: consta de dos o més elements.
- *Esdeveniment impossible*: és el que mai pot realitzar-se (està determinat pel conjunt buit, $\varnothing$).
- *Esdeveniment segur*: és el que sempre es compleix (està determinat pel conjunt total, E).
- *Esdeveniments disjunts o mútuament excloents*: aquells esdeveniments A i B que no poden realitzar-se conjuntament, $A \cap B = \varnothing$.

Aclarim aquests conceptes amb un exemple:

Exemple 3

Es realitza l'experiment aleatori «llançar un dau»:

- Espai mostral: $E = \{1, 2, 3, 4, 5, 6\}$
- Esdeveniment elemental: traure un 2 $= \{2\}$
- Esdeveniment compost: traure un nombre imparell $= \{1, 3, 5\}$
- Esdeveniment impossible: traure un 7 $= \{\varnothing\}$

- Esdeveniment segur: traure un nombre més petit que $7 = \{1, 2, 3, 4, 5, 6\} = E$
- Esdeveniments disjunts: A = traure un nombre parell = $\{2, 4, 6\}$
  B = traure un nombre senar = $\{1, 3, 5\}$

Àlgebra de successos és el conjunt format per tots els esdeveniments. Se sol representar per $\wp(E)$.

Exemple 4

Es considera l'experiment «llançar una moneda» i es calcula l'àlgebra d'esdeveniments de l'experiment:

Nombre d'esdeveniments elementals		Esdeveniments que el formen		
0	elements	$\{\varnothing\}$	——>	1 subconjunt
1	element	$\{C\}, \{X\}$	——>	2 subconjunts
2	elements	$\{C, X\} = E$	——>	1 subconjunt

Total, 4 subconjunts ($4 = 2^2$) i $\wp(E) = \{\{\varnothing\}; \{C\}; \{X\}; \{C, X\}\}$

Nota

Tenint en compte que els esdeveniments són subconjunts de E, es pot aplicar la teoria general de conjunts: unions, les interseccions, diferències i complementaris...

6.3.2. Operacions amb esdeveniments

En tot moment es considerarà E = espai mostral.

Es considerarà l'experiment següent per a mostrar els exemples:

> Experiment «llançar una trompa amb 4 possibles resultats»: {1, 2, 3, 4}
> Els esdeveniments elementals són:
> {Eixir un 1}, {Eixir un 2}, {Eixir un 3}, {Eixir un 4}

No obstant això, abans de començar amb la definició de les diferents operacions, és necessari aclarir els conceptes *igualtat* i *inclusió*.

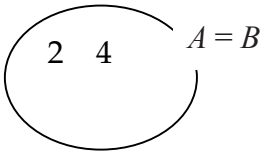
Igualtat

Es diu que els esdeveniments A i B són iguals ($A = B$) quan l'ocurrència de A implica l'ocurrència de B, i l'ocurrència de B implica l'ocurrència de A. És a dir, tots dos esdeveniments estan formats pels mateixos esdeveniments elementals.

Exemple 5

Si $A = \{2, 4\}$ i $B = \{\text{Eixir parell}\} \rightarrow$ els esdeveniments A i B són iguals.

Gràficament:



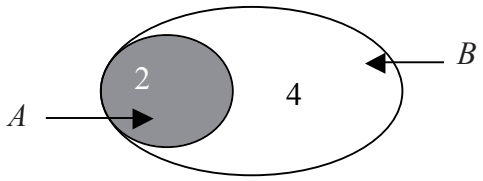
Inclusió

Es diu que l'esdeveniment A està inclòs en l'esdeveniment B ($A \subseteq B$) quan l'ocurrència de A implica l'ocurrència de B. És a dir, els esdeveniments elementals de A, estan tots també en B.

Exemple 6

Si $A = \{\text{Eixir 2}\}$ i $B = \{\text{Eixir parell}\}$, A implica B

Gràficament:



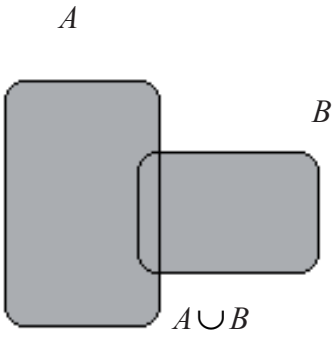
Nota

Cal notar que en aquest cas B no implica A: encara que isca parell, no ha d'eixir necessàriament 2.

Unió d'esdeveniments: ($\cup$)

Donats dos esdeveniments qualssevol A i B, l'esdeveniment $A \cup B$ (A unió B) és un succés que es verifica quan es verifica A o es verifica B, o quan es verifiquen tots dos al mateix temps.

Exemple 7



Es consideren els esdeveniments compostos A = {Eixir parell} i B = {Eixir més alt que 2} Així A = {2, 4} i B = {3, 4} i, per tant,

$$A \cup B = \{2, 3, 4\}$$

És a dir, $A \cup B$ és l'esdeveniment que ocorre quan el resultat de l'experiment és 2, 3 o 4.

Propietats de la unió

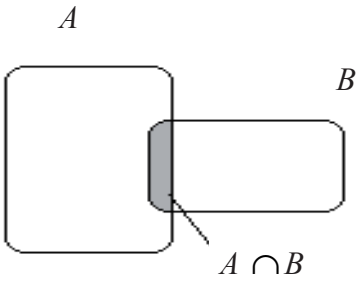
Siguen A, B i C tres esdeveniments qualssevol de l'experiment, llavors es compleix:

$$A \cup (B \cup C) = (A \cup B) \cup C$$
$$A \cup B = B \cup A$$
$$A \cup E = E$$
$$A \cup \varnothing = A$$

Intersecció d'esdeveniments ($\cap$)

Donats dos esdeveniments qualssevol A i B, l'esdeveniment $A \cap B$ (A intersecció B) és un esdeveniment que es verifica quan ocorren A i B simultàniament.

Exemple 8



Es consideren els esdeveniments compostos A ={Eixir parell} i B = {Eixir més alt que 2} Així A = {2,4} i B = {3,4} i, per tant,

$$A \cap B = \{4\}$$

És a dir, $A \cap B$ és l'esdeveniment que ocorre quan el resultat de l'experiment és 4.

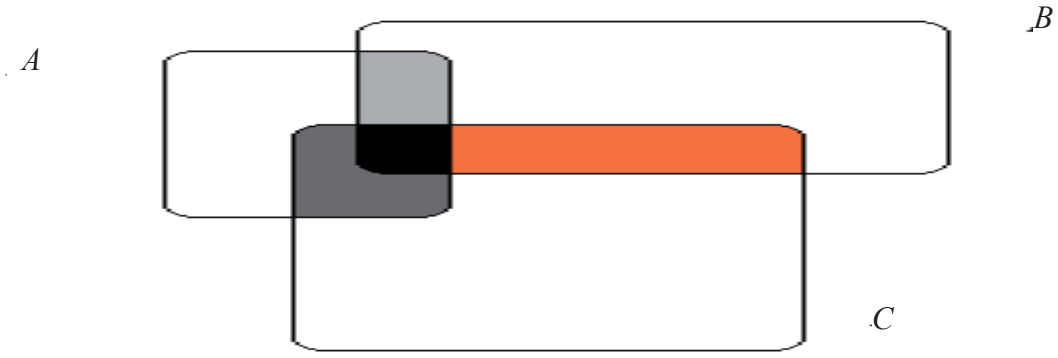
Propietats de la intersecció

Siguen A, B i C tres esdeveniments qualssevol de l'experiment, llavors es compleix que:

$$A \cap (B \cap C) = (A \cap B) \cap C$$
$$A \cap B = B \cap A$$
$$A \cap E = A$$
$$A \cap \varnothing = \varnothing$$

Propietats distributives de la unió i de la intersecció

Es consideren tres esdeveniments qualssevol A, B i C, llavors es compleix que:

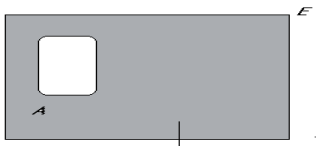


- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

Esdeveniment negació o complementari

Donat un esdeveniment A, s'anomena *esdeveniment negació* o *esdeveniment complementari* de A, es denota per $\overline{A}$, l'esdeveniment que ocorre quan –i només quan– no succeix A.

Exemple 9



La zona ombrejada és l'esdeveniment complementari de l'esdeveniment A

Es considera l'esdeveniment
A = {Eixir parell},
llavors $\overline{A}$ = {1, 3}

Propietats de l'esdeveniment complementari

a) $\overline{\varnothing} = E$ i $\overline{E} = \varnothing$
b) $A \cup \overline{A} = E$ i $A \cap \overline{A} = \varnothing$
c) $\overline{\overline{A}} = A$
d) $\overline{A \cup B} = \overline{A} \cap \overline{B}$ i $\overline{A \cap B} = \overline{A} \cup \overline{B}$ (Lleis de De Morgan)

Esdeveniments independents

Dos esdeveniments són independents quan l'ocurrència d'un no influeix en l'ocurrència de l'altre.

Exemple 10

Els esdeveniments *A* (llançar una moneda) i *B* (llançar un dau) són independents. El resultat de l'un no influeix en el resultat de l'altre.

Els resultats d'extraure dues cartes d'una baralla són independents? És a dir, el resultat de la primera carta afecta el resultat de la segona? Depèn de si tornem o no a la baralla la carta extreta. Diguem-ne que l'experiment pot ser amb reemplaçament o no. Si la carta es torna a la baralla els esdeveniments seran independents i, en un altre cas, no ho seran.

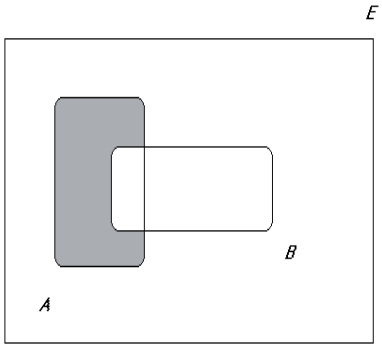
Nota

Dos esdeveniments disjunts (aquells que no tenen cap element en comú) no són independents, ja que si un ocorre, és segur que l'altre no ho farà. Per tant, l'ocurrència de l'un influeix en l'ocurrència de l'altre.

Diferència de dos esdeveniments

Donats els esdeveniments *A* i *B*, s'anomena *esdeveniment A menys B* i es denota $A - B$, aquell en què es verifica *A* però no *B*, és a dir, $A - B$ està format pels esdeveniments elementals que estan en *A* i no estan en *B*. És evident que $A - B = A \cap \overline{B}$.

Exemple 11



A = {Eixir més alt que 1} = {2, 3, 4}
B = {Eixir parell} = {2, 4}
$A - B$ = {3}

Exemple 12

En una empresa ubicada a Castelló tots els treballadors tenen almenys una titulació universitària, alguns dels treballadors són llicenciats i uns altres són diplomats. Per altra part, hi ha un grup que treballa habitualment a l'oficina central de Castelló, uns altres treballen des de casa i un grup reduït ho fa a l'estranger.

Es consideren els esdeveniments següents:

A: Ser llicenciat
B: Ser diplomat
C: Treballar a l'oficina central de Castelló
D: Treballar des de casa
E: Treballar a l'estranger

a) Identifica i explica els esdeveniments següents: $A \cup B$; $A \cup E \cup C$; $A \cap D$; $D \cap E$; $\overline{A}$; $A - E$; $A \cap (E \cup C)$; $(A \cap E) \cup (A \cap C)$.
b) Quina diferència hi ha entre $A \cup B$ i $A \cap B$?
c) És possible que el 50% dels treballadors siguen llicenciats i el 65% diplomats?
d) Si l'empresa té 200 treballadors, 100 són llicenciats i 125 són diplomats, assigna intuïtivament la probabilitat d'escollir una persona a l'atzar de l'empresa que:

i. siga diplomada
ii. siga llicenciada
iii. siga diplomada o llicenciada
iv. siga diplomada i llicenciada.

Les respostes són:

a)

$A \cup B$ = Són tots aquells treballadors de l'empresa llicenciats o diplomats. En aquest cas, com que tots els alumnes són diplomats o llicenciats, la unió dels dos conjunts és el conjunt de tots els treballadors de l'empresa.

$A \cup E \cup C$	= Són tots aquells treballadors de l'empresa llicenciats o que treballen a l'estranger o a l'oficina.
$A \cap D$	= Són aquells treballadors llicenciats i que treballen des de casa.
$D \cap E$	= Són els treballadors que treballen des de casa i que treballen a l'estranger. En aquest cas serien les persones que viuen a l'estranger i que treballen per a l'empresa.
$\overline{A}$	= Són tots aquells treballadors de l'empresa que no són llicenciats.
$A - E$	= Són aquells treballadors llicenciats que no treballen a l'estranger.
$A \cup (E \cup C)$	= Són els treballadors llicenciats que treballen a l'oficina central de Castelló o que treballen a l'estranger.
$(A \cap E) \cup (A \cap C)$	= Són aquells treballadors llicenciats que treballen a l'estranger, o els llicenciats que treballen a l'oficina central de Castelló.

b)

$A \cap B$ és el conjunt format per aquells treballadors llicenciats i diplomats. Aquest està inclòs en $A \cup B$, ja que en aquest conjunt hi són tots, és a dir, aquells que són únicament llicenciats, els que són únicament diplomats i aquells que tenen totes dues titulacions.

c)

Sí que és possible, és el cas en què hi ha persones amb dues titulacions.

d)

P (ser diplomada) = $\frac{100}{200} = \frac{1}{2}$ P (ser llicenciada) = $\frac{125}{200} = \frac{5}{8}$.

P (diplomada o llicenciada)= $\frac{200}{200} = 1$, ja que 100 + 125 > 200.

P (diplomada i llicenciada)= $\frac{25}{200} = \frac{1}{8}$, ja que 100 + 125 = 225 i, per tant, entre llicenciats i diplomats hi ha 225 persones. Com que en total hi ha 200 treballadors, hi ha 25 persones que s'han comptat dues vegades i, en conseqüència, tenen totes dues titulacions.

Exemple 13

En una empresa de Castelló treballen 200 persones. 80 saben parlar anglès, 110 francès i 60 alemany. 50 persones saben parlar francès i anglès, 35 saben parlar francès i alemany i 40 saben parlar anglès i alemany. A més a més, 30 treballadors saben parlar tots tres idiomes.

En aquestes condicions, contesta a les pregutes següents:

1. Quantes persones saben parlar únicament anglès?
2. Calcula la probabilitat que en escollir a l'atzar una persona de l'empresa, únicament sàpia parlar anglès.
3. Calcula la probabilitat que en escollir a l'atzar una persona de l'empresa, no sàpia parlar cap dels tres idiomes.

1.

Si emprem la notació A = Anglès; F = Francès; i Al = Alemany, es té que:

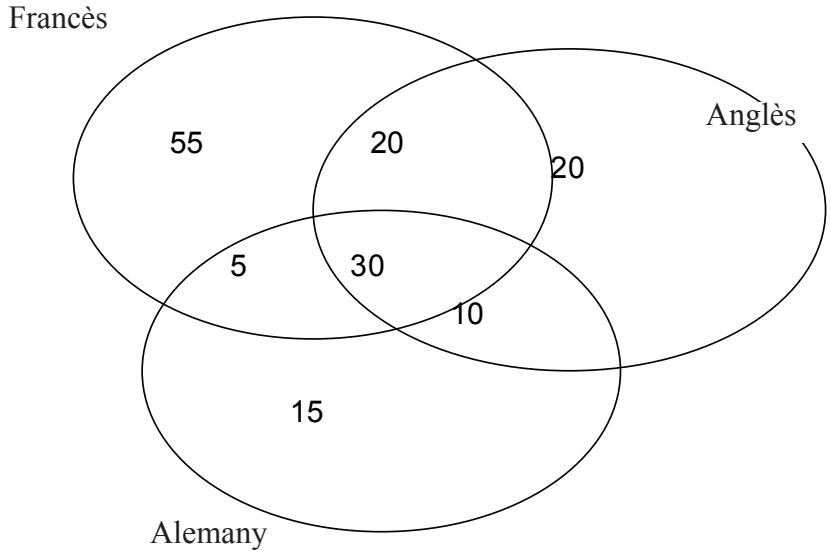
- $A \cap Al \cap F$ són 30.
- $A \cap Al$ són 40.
- $Al \cap F$ són 35.
- $A \cap F$ són 50.
- En conseqüència, i per no comptar dues vegades les mateixes persones:

  El nombre de persones que saben únicament francès són 110 – 35 – 50 + 30 = 55. (És a dir, al total de persones que saben francès, cal restar-li aquelles que també parlen alemany i francès (35) i també les que parlen anglès i francès (50). No obstant això, per a no restar dues vegades aquelles persones que saben tots tres idiomes, és necessari sumar-li una vegada les persones que els parlen, tots tres (30)).

  - El nombre de persones que saben únicament anglès són 80 – 50 – 40 + + 30 = 20. (La raó és semblant a la del cas anterior.)

  - El nombre de persones que saben únicament alemany són 60 – 35 – 40 + + 30 = 15. (La raó és semblant a la del cas anterior.)

Aquesta informació es pot aclarir molt més mitjançant el gràfic següent:



2.

P (sàpia parlar únicament francès) $= \frac{60}{200} = \frac{3}{10}$.

3.

El nombre de persones que no saben parlar cap dels tres idiomes és:

$200 - (60 + 5 + 20 + 30 + 20 + 10 + 15) = 155$.

P (que no sàpia cap idioma) $= \frac{45}{200} = \frac{9}{40}$.

6.4. Concepte de *probabilitat*. Definició i propietats

Quan es va introduir el concepte *probabilitat* en els primers punts del tema, se'n distingien dues interpretacions: la basada en la freqüència, en la qual es definia la probabilitat d'un esdeveniment com el valor al qual s'acosta la freqüència relativa d'aquest esdeveniment en realitzar l'experiment un nombre molt gran de vegades; i la subjectiva, en què la probabilitat d'un esdeveniment era assignada segons la informació que es coneixia de l'experiment.

És evident que els conceptes de *probabilitat* més formal o clàssic i el basat en la freqüència han de coincidir, perquè la probabilitat d'un esdeveniment ha de ser un únic valor.

És el moment ara de formalitzar-ne el concepte.

Per a fer-ho, es definiran els axiomes que determinen el concepte clàssic de *probabilitat*, basant-se en 3 propietats fonamentals que compleix la probabilitat definida per les freqüències.

Així, es considera E l'espai mostral d'un experiment aleatori, i A i B dos esdeveniments qualssevol de l'experiment:

Cal recordar que:

$P(A)$ = freqüència relativa de A =

$$fr(A) = \frac{\text{nre. d'ocurrències } A}{\text{nre. de vegades que s'ha realitzat l'experiment}} = \frac{n_A}{n}$$

quan n tendeix a l'infinit.

Es denoten:

n = nombre de vegades que s'ha realitzat l'experiment
n_A = nombre de vegades que ha ocorregut A
n_B = nombre de vegades que ha ocorregut B

Amb aquesta notació, passen a comentar les tres propietats:

Propietat 11

Per a qualsevol esdeveniment A, es té que $0 \leq fr\,(A) \leq 1$.

És evident, doncs, que $0 \leq n_A$ i $n > 0$. Per això $0 \leq \frac{n_A}{n}$. Per altre costat, és evident que $n_A \leq n$ i, consegüentment, $\frac{n_A}{n} \leq 1$.

Unint totes dues desigualtats, s'obté que $0 \leq \frac{n_A}{n} \leq 1$, que demostra la propietat.

Propietat 2

$fr(E) = 1$. És evident, ja que E és l'esdeveniment segur, per ser E l'espai mostral.

Propietat 3

Si A i B són dos esdeveniments disjunts, $A \cap B = \varnothing$, aleshores es compleix:

$fr(A \cup B) = fr(A) + fr(B)$.

$$fr(A \cup B) = \frac{\text{nombre de vegades que ocorre } A \text{ o } B}{\text{nombre de vegades que es realitza l'experiment}} =$$

$$= \frac{\text{nombre de vegades de } A + \text{nombre de vegades de } B - \text{nombre de vegades de tots dos}}{\text{nombre de vegades que es realitza l'experiment}}$$

$$= \frac{n_A + n_B - 0}{n} = \frac{n_A}{n} + \frac{n_B}{n} = fr(A) + fr(B)$$

Basant-se en aquestes propietats, es pot definir ja el concepte més formal i clàssic de *probabilitat*.

6.4.1. Definició clàssica o axiomàtica de *probabilitat*

Donat E l'espai mostral d'un experiment aleatori, una probabilitat en E és qualsevol funció que assigne a cada esdeveniment A de l'experiment, un nombre $P(A)$, de tal manera que es complisquen els axiomes següents:

Axioma 1: la probabilitat de l'esdeveniment segur val 1. $P(E) = 1$.
Axioma 2: la probabilitat de qualsevol altre esdeveniment A és no negativa: $P(A) \geq 0$.
Axioma 3: la probabilitat de la unió de dos esdeveniments mútuament excloents, A i B, és la suma de les seues probabilitats.
És a dir, si $A \cap B = \varnothing$ llavors $P(A \cup B) = P(A) + P(B)$.

Aquest darrer axioma pot generalitzar-se de la manera següent: la probabilitat de la unió d'un conjunt d'esdeveniments mútuament excloents és igual a la suma de les seues probabilitats:

$$P(\cup A_i) = \sum P(A_i) = P(A_1) + P(A_2) + ... + ...$$

Propietats de la probabilitat

D'aquests axiomes es poden deduir una sèrie de propietats:

Propietat 1

Si $A_1, A_2, ..., A_n$ són esdeveniments disjunts dos a dos amb $n > 2$ (o siga, $A_i \cap A_j = \varnothing$ amb $i \neq j$), llavors: $P(A_1 \cup A_2 \cup ... \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n)$. A més a més, si $B = A_1 \cup A_2 \cup ... \cup A_i$ $P(B) = P(A_1) + P(A_2) + ... + P(A_i)$.

És immediata per l'axioma 3, ja que el nombre d'esdeveniments que s'han pres és n (un nombre finit), i com que la propietat es compleix per a dos esdeveniments i per a una quantitat infinita numerable, es compleix per a una quantitat finita.

Exemple 14

Els resultats del baròmetre del CIS al juliol de 2006 a la pregunta «Per a començar, referint-nos a la situació econòmica general d'Espanya, com la qualificaria vostè: de molt bona, bona, regular, roïna o molt roïna?», han estat els que mostra la taula:

Molt bona	Bona	Regular	Roïna	Molt roïna	NS	NC	TOTAL
0,8	21,0	49,0	20,7	7,3	1,2	0,2	100,0

Calcula la probabilitat que en escollir a l'atzar una persona de les que han contestat el qüestionari, la seua resposta haja estat «Molt bona», «Bona», «Regular» o «Molt roïna». Calcula també la probabilitat que una persona opine que la pel·lícula ha sigut regular o millor.

Com que les dades apareixen en percentatges, es dedueixen fàcilment les probabilitats dels esdeveniments elementals:

- P (Contestar «Molt bona») = 0,008
- P (Contestar «Regular») = 0,49
- P (Contestar «Molt roïna») = 0,073
- P (NC) = 0,002.

P (Contestar «Bona») = 0,21
P («Roïna») = 0,207
P (NS) = 0,01

Aleshores P (Contestar «Molt bona» o «Bona» o «Regular» o «Molt roïna») =

= P(Molt bona $\cup$ Bona $\cup$ Regular $\cup$ Molt roïna) = P(Molt bona) + P(Bona) + + P(Regular) + P(Molt roïna) = 0,008 + 0,21 + 0,49 + 0,073 = 0,781.

Ja que els esdeveniments són tots disjunts (una persona no pot haver contestat dues respostes diferents a aquesta pregunta).

D'altra banda, com que l'esdeveniment regular o millor = Molt bona $\cup$ Bona $\cup$ Regular, llavors, P(regular o millor) = P(Molt bona $\cup$ Bona $\cup$ Regular) = = P(Molt bona) + P(bona) + P(regular) = 0,008 + 0,21 + 0,49 = 0,708.

Propietat 2

$P(\overline{A}) = 1 - P(A)$, on A és un esdeveniment qualsevol. (Nota: $\overline{A}$ és l'esdeveniment complementari de A.)

$A \cup \overline{A} = E \longrightarrow P(A \cup \overline{A}) = P(E) = 1$

I com que $A \cap \overline{A} = \varnothing$ ---Axioma 3 $\longrightarrow P(A \cup \overline{A}) = P(A) + P(\overline{A})$

D'ambdues conseqüències, $P(A) + P(\overline{A}) = 1 \longrightarrow P(\overline{A}) = 1 - P(A)$.

Exemple 15

El 8% de les persones que viuen en una ciutat són estrangeres europees, el 24% són estrangeres no europees i la resta són nascudes a Espanya. Calcula la probabilitat que en escollir a l'atzar una persona d'aquesta ciutat siga de nacionalitat espanyola.

D'una banda es té que:

P (ser estranger europeu) = 0,08 i P (ser estranger no europeu) = 0, 24. Per tant:

P (ser espanyol) = 1 – P (ser estranger europeu $\cup$ ser estranger no europeu) =

= 1 – (P (ser estranger europeu) + P (ser estranger no europeu)) = 1 – 0,32 = 0,68.

Ja que l'esdeveniment complementari de (ser estranger europeu $\cup$ ser estranger no europeu) és ser espanyol.

Propietat 3

$P(\varnothing) = 0$

$\varnothing = \overline{E} \longrightarrow P(\varnothing) = P(\overline{E})$

Per la propietat 2, $P(\overline{E}) = 1 - P(E) = 1 - 1 = 0$.

Per tant, $P(\varnothing) = 0$.

Exemple 16

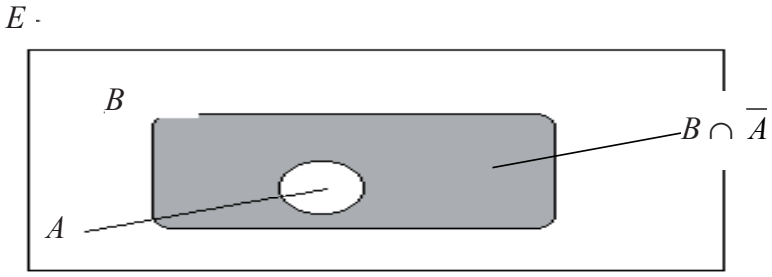
En una inversió borsària de trenta-quatre valors, s'hi han produït pèrdues en tots. Calcula la probabilitat d'obtenir-hi cap guany.

En aquest cas, $\varnothing$ és l'esdeveniment «obtenir cap guany», ja que segons la informació del problema aquest esdeveniment és impossible. Per tant P(obtenir cap guany) = 0.

Propietat 4

Donats dos esdeveniments A i B en què $A \subseteq B \Rightarrow P(A) \leq P(B)$.

Observant el gràfic es dedueix que $B = A \cup (B \cap \overline{A})$:



A més, $A \cap (B \cap \overline{A}) = \varnothing \longrightarrow A$ i $(B \cap \overline{A})$ són disjunts.

Per tant, per l'axioma 3: $P(B) = P(A) + P(B \cap \overline{A})$.

Com que per l'axioma 1, $P(B \cap \overline{A}) \geq 0 \longrightarrow P(B) \geq P(A)$.

Exemple 17

Una empresa té 200 treballadors, 150 dels quals treballen a les oficines i la resta són operaris amb tasques diferents. Dels que treballen a les oficines, 29 tenen algun tipus de càrrec i la resta són administratius i comercials. Si s'escull a l'atzar una persona que treballe a l'empresa, calcula la probabilitat que aquesta tinga un càrrec. Calcula també la probabilitat que treballe a les oficines. Quina de les dues probabilitats és més alta?

Atenent les dades:

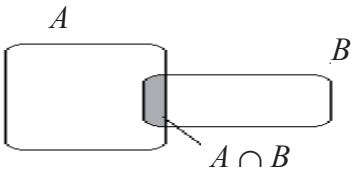
$$P(\text{tenir un càrrec}) = \frac{29}{200} \quad \text{i} \quad P(\text{treballar a les oficines}) = \frac{150}{200}.$$

En conseqüència, P(tenir un càrrec) < P(treballar a les oficines), la qual cosa és coherent amb la propietat 4, ja que l'esdeveniment «tenir un càrrec» està inclòs en «treballar a les oficines» (totes les persones que tenen un càrrec treballen a les oficines).

Propietat 5

Per a tot A, B que siguen de E, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Intuïtivament:



En fer $A \cup B$, es pren dues vegades $A \cap B$, després per a calcular el que es vol s'ha de restar una vegada $A \cap B$.

Amb més formalitat, tenint present que:

$$A \cup B = A \cup (B \cap \overline{A}) \text{ i que } B = (A \cap B) \cup (B \cap \overline{A}),$$

s'obté, d'una banda, $P(A \cup B) = P(A \cup (B \cap \overline{A}) = P(A) + P(B \cap \overline{A})$

i, d'altra banda, $P(B) = P(A \cap B) \cup (B \cap \overline{A}) = P(A \cap B) + P(B \cap \overline{A})$.

Restant ambdues expressions, s'obté $P(A \cup B) - P(B) = P(A) - P(B \cap A)$.

Exemple 18

En una selecció de personal per a un treball administratiu es pregunta als candidats si saben fer presentacions. Les respostes han estat les següents: el 60% sap realitzar-les emprant el software, el 75% les sap fer emprant transparències i hi ha un 15% que contesta que no sap fer-ne.

Calcula la probabilitat que un candidat escollit a l'atzar sàpia fer presentacions amb software i amb transparències.

Es demana P(saber fer presentacions software $\cup$ saber fer presentacions amb transparències).

Per altra part, segons l'enunciat:

- P(saber fer presentacions software) = $P(sfps)$ = 0,6
- P(saber fer presentacions amb transparències) = $P(sfpt)$ = 0,75

$$P(\text{no saber fer presentacions}) = P(nsfp) = 0{,}15$$

D'altra banda, $P(sfps \cup sfpt) = 1 - P(nsfp) = 1 - 0{,}15 = 0{,}85$.

Però segons la propietat:

$$P(sfps \cup sfprt) = P(sfps) + P(sfpt) - P(sfps \cap sfpt)$$

Substituint,

$$0{,}85 = 0{,}6 + 0{,}75 - P(sfps \cap sfpt) \qquad P(sfps \cap sfpt) = 0{,}5$$

Corol·lari

Donats dos esdeveniments A i B en què $B \subseteq A \qquad P(A - B) = P(A) - P(B)$.

Tenint en compte que $P(A - B) = P(A \cap \overline{B})$ i la propietat 5, s'obté que:

$$P(A \cap \overline{B}) = P(A) + P(\overline{B}) - P(A \cup \overline{B}).$$

I com que $B \subseteq A \ \overline{A} \subseteq \overline{B}$, llavors l'espai mostral E compleix:

$$E = A \cup \overline{A} \subseteq A \cup \overline{B} \subseteq E \qquad A \cup \overline{B} = E \rightarrow P(A \cup \overline{B}) = 1.$$

$$P(A \cap \overline{B}) = P(A) + P(\overline{B}) - P(A \cup \overline{B}) = P(A) + (1 - P(B)) - 1.$$

Es té que: $P(A - B) = P(A) - P(B)$.

Exemple 19

En una empresa treballen un 25% de persones estrangeres. Se sap que hi ha únicament un 8% de treballadors que tenen un nivell B2 d'anglès i tots són estrangers. S'escull una persona a l'atzar per a fer un estudi estadístic sobre el nivell de coneixements en llengües de les empreses espanyoles, quina és la probabilitat que s'esculla una persona estrangera que no tinga el nivell B2 d'anglès?

Si es denoten *Es = escollir una persona estrangera* i *B = escollir una persona que tinga el nivell B2 d'anglès* el que es demana és $P(Es \cap \overline{B}) = P(Es - B)$. D'altra banda com que $B \subseteq Es$ $P(Es - B) = P(Es) - P(B) = 0{,}25 - 0{,}08 = 0{,}17$.

6.4.2. Espais mostrals finits. Regla de Laplace

S'anomenen *espais mostrals finits* els espais mostrals que provenen d'experiments per als quals només hi ha un nombre finit de resultats possibles.
Així: $E = \{w_1, w_2, \ldots, w_n\}$.

En un experiment aleatori amb un espai mostral finit, una distribució de probabilitat s'especifica assignant una probabilitat p_i a cada resultat w_i que pertany a E, $p_i = P(\{w_i\})$. Ha de complir-se que:

a) $p_i \geq 0$

b) $P(E) = 1 \longrightarrow \sum_{i=1}^{n} p_i = p_1 + \ldots + p_n = 1$

Ja que els esdeveniments $\{w_1, w_2, \ldots, w_n\}$ són elementals i, per tant, disjunts.

En aquestes condicions, si $A = \{w_i \ldots, w_{ij}\}$, es té $P(A) = p_i + \ldots + p_{ij}$.

S'anomenen *espais mostrals simples* els espais mostrals finits en què tots els resultats són equiprobables (tenen la mateixa probabilitat). Si $E = \{w_1, w_2, \ldots, w_n\}$, llavors:

$$P(\{w_i\}) = \frac{1}{n},\ i = 1, \ldots, n.$$

En aquests espais mostrals simples, donat un esdeveniment $A = \{w_1, w_2, w_k\}$ amb $k < n$ es té que:

$$P(A) = P(w_1 \cup w_2 \cup \ldots\ldots \cup w_k) = P(w_1) + P(w_2) + \ldots + P(w_k)$$

$$= \frac{1}{n} + \frac{1}{n} + \ldots\ldots + \frac{1}{n} = k \cdot \frac{1}{n} = \frac{\text{nre. d'esdeveniments elementals de } A}{\text{nre. d'esdeveniments elementals de } E}$$

Aquest fet justifica la fórmula de Laplace:

Donat un experiment aleatori on tots els resultats (esdeveniments elementals) tenen la mateixa probabilitat de produir-se, si A és un esdeveniment possible de l'experiment es compleix que:

$$P(A) = \frac{\text{Casos favorables a } A}{\text{Casos possibles}}$$

On els casos possibles són tots els resultats possibles de l'experiment (el nombre total d'esdeveniments elementals) i els casos favorables a A és el nombre de resultats que compleixen A (nombre d'esdeveniments elementals que compleixen A).

Exemple 20

Si es llança una dau sense trucar, quina és la probabilitat d'obtenir una xifra senar en el resultat?

És evident que l'espai mostral corresponent és $E = \{1, 2, 3, 4, 5, 6\}$ i que tots els resultats de l'experiment tenen la mateixa probabilitat de produir-se.

Siga l'esdeveniment A = eixir un resultat senar = $\{1, 3, 5\}$.

Sense emprar-hi Laplace

Així, $P(\text{eixir 1}) = P(\text{eixir 2}) = P(\text{eixir 3}) = P(\text{eixir 4}) = P(\text{eixir 5}) = P(\text{eixir 6}) = \frac{1}{6}$.

$P(A) = P(\text{eixir 1} \cup \text{eixir 2} \cup \text{eixir 3}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$.

Emprant-hi Laplace

Casos possibles: 6; ja que l'espai mostral té 6 esdeveniments elementals.

Casos favorables: 3; ja que dels resultats possibles únicament 3 compleixen ser senars (1, 3, 5).

Aplicant-hi Laplace $\rightarrow P(A) = \frac{\text{Casos favorables a } A}{\text{Casos possibles}} = \frac{3}{6} = \frac{1}{2}$.

Exemple 21

Un estudiant decideix presentar-se a unes oposicions a l'administració educativa en acabar la carrera universitària. La primera fase de l'oposició consisteix a desenvolupar un tema de deu possibles. Si l'estudiant únicament sap els temes parells:

a) Descriu l'espai mostral de l'experiment «eixir a l'atzar un tema dels deu». És finit?
b) Calcula la probabilitat de cada esdeveniment elemental i comprova que si se sumen el resultat és 1.
c) Calcula la probabilitat que isca un tema dels que sap l'estudiant.

a)

En aquest cas l'espai mostral és molt senzill: $E = \{$eixir el tema 1, eixir el tema 2, eixir el tema 3, eixir el tema 4, eixir el tema 5, eixir el tema 6, eixir el tema 7, eixir el tema 8, eixir el tema 9, eixir el tema 10$\}$.

És evidentment finit, ja que l'espai mostral té 10 esdeveniments elementals.

b)

Com que cada esdeveniment té la mateixa probabilitat de produir-se, es compleix que:

$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = P(7) = P(8) = P(9) = P(10) = \frac{1}{10}$.

On 1 = eixir el tema 1; 2 = eixir el tema 2; ...

A més a més, $P(1) + P(2) + P(3) + P(4) + P(5) + P(6) + P(7) + P(8) + P(9) + P(10)$

$$= \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} = 10 \cdot \frac{1}{10} = 1.$$

c)

En aquest cas s'aplicarà la regla de Laplace.

S'anomena A l'esdeveniment «temes que s'ha estudiat l'alumne». Així $A = \{1, 3, 5, 7, 9\}$.

Per tant:

Casos possibles: 10 (ja que hi ha 10 resultats possibles de l'experiment)

Casos favorables a A: 5 (ja que l'estudiant ha estudiat 5 temes)

Per tant, aplicant-hi Laplace es té que: $P(A) = \frac{\text{Casos favorables a } A}{\text{Casos possibles}} = \frac{5}{10} = \frac{1}{2}$

Nota

Cal remarcar que per a poder aplicar la regla de Laplace és necessari que tots els esdeveniments de l'espai mostral tinguen la mateixa probabilitat de succeir.

Exemple 22

Es llança una moneda dues vegades i es compta el nombre de cares que ixen. Quina és la probabilitat que isca una cara?

L'experiment consisteix a llançar dues vegades una moneda i comptar el nombre de cares que ixen. Així, l'espai mostral corresponent és $E =$ {obtenir 0 cares, obtenir 1 cara, obtenir 2 cares}.

Tots els esdeveniments elementals d'aquest experiment tenen la mateixa probabilitat de produir-se? Una anàlisi ràpida faria pensar que sí i llavors tots tindrien una probabilitat $\frac{1}{3}$, però una anàlisi més exhaustiva respon negativament a la pregunta, ja que obtenir una cara és més probable que obtenir dues cares o que no obtenir-ne cap. Per què?

Per a respondre a la pregunta cal observar els resultats de l'experiment «llançar dues vegades una moneda», que precedeix el recompte del nombre de cares. Els resultats d'aquest experiment són: $\{(C, C), (C, +), (+, C), (+, +)\}$ (on C = cara i + = creu).

En aquest cas tots els resultats tenen la mateixa probabilitat de produir-se, ja que el resultat de cada llançament és completament independent dels resultats dels altres.

Per tant, $P(C, C) = P(C, +) = P(+, C) = P(+, +) = \frac{1}{4}$, i, en conseqüència:

$P(\text{obtenir una cara}) = P((C, +) \cup (+, C)) = P(C, +) + P(+, C) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.

6.5. Teorema de la probabilitat total. Teorema de Bayes

Abans de definir el teorema de la probabilitat total, és necessari estudiar la probabilitat d'un esdeveniment condicionada a l'ocurrència d'un altre.

6.5.1. Probabilitat condicionada

En una carrera de cavalls en què en participen 15, un jugador ha apostat pel cavall número 2. El jugador rep dues hores després de l'aposta la informació que el nombre del cavall guanyador és parell. Per altra part, un segon jugador rep la mateixa informació abans d'apostar i decideix fer-ho també pel cavall número 2.

Si tots dos han decidit el número del cavall aleatòriament, quin dels dos jugadors té més probabilitat de guanyar?

És evident que la informació de la qual gaudeix el segon jugador l'afavoreix a l'hora de prendre la decisió.

1r jugador

Segons diu l'enunciat, a priori tots els cavalls tenen la mateixa probabilitat de guanyar i, en conseqüència, es pot aplicar la regla de Laplace a l'hora de calcular la probabilitat d'encert del cavall guanyador. Així:

Casos possibles: 15 (ja que són 15 els possibles cavalls guanyadors)

Casos favorables: 1 $P(\text{guanyar jug. 1}) = P(\text{guanyar el cavall 2}) = \frac{1}{15}$

2n jugador

El segon jugador gaudeix d'una informació privilegiada abans d'apostar. Sap que el cavall guanyador duu un nombre parell. Aplicant-hi la regla de Laplace:

> Casos possibles: 7 (ja que són 7 els possibles cavalls guanyadors, perquè, per la informació rebuda, els possibles guanyadors són els cavalls 2, 4, 6, 8, 10, 12, 14).
>
> Casos favorables: 1 P(que guanye el jugador 2) = P(que guanye el cavall 2) = $\frac{1}{7}$.

Així doncs, el jugador 2 té, abans de començar la carrera, més probabilitat d'encertar el cavall guanyador que el jugador 1.

Aquest exemple posa de manifest la importància de conèixer informació sobre els resultats de l'experiment a l'hora d'establir les probabilitats. En aquest exemple, si anomenem A l'esdeveniment «que guanye el cavall núm. 2» i B l'esdevenient «que guanye un cavall amb nombre parell», la probabilitat de l'esdeveniment A no és la mateixa si se sap que passa B que si no se sap.

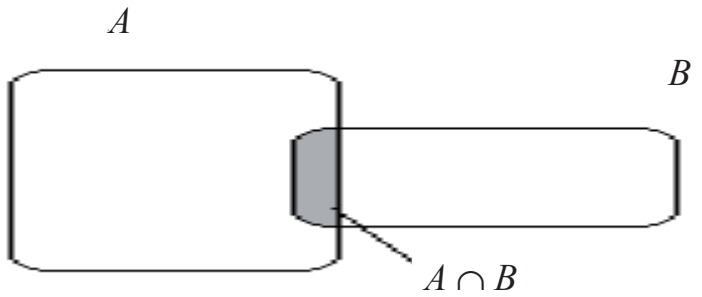
D'aquesta manera, en el primer cas $P(A) = \frac{1}{15}$,

i en el segon cas $P(A$ sabent que passa $B) = P(A/B) = \frac{1}{7}$.

El càlcul de probabilitats en aquest tipus de situacions s'anomena *càlcul de probabilitats condicionades*, perquè l'ocurrència o no d'uns esdeveniments influeix en les probabilitats dels altres.

Definició de *probabilitat condicionada*

Donats dos esdeveniments A i B de l'espai mostral E, s'anomena *probabilitat de A condicionada a B* i s'escriu $P(A/B)$ la probabilitat que ocórrega l'esdeveniment A considerant que abans ha ocorregut l'esdeveniment B.



Si ha ocorregut B, es tindrà un nou espai mostral, $E_B = E \cap B$, ja que l'ocurrència de qualsevol altre esdeveniment ha de tenir en compte que B ha ocorregut, i així:

$$\mathrm{P}(A/B) = \frac{\text{nombre casos favorables en } A \cap B}{\text{nombre casos favorables en } B} = \frac{\dfrac{\text{nre. casos favorables } A \cap B}{\text{nre. casos possibles en } E}}{\dfrac{\text{nre. casos posibles en } B}{\text{nre. casos possibles en } E}}$$

$$= \frac{P(A \cap B)}{P(B)}.$$

És a dir:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}.$$

Nota

En ocasions, la fórmula $P(\mathrm{A}/B)$ trobada anteriorment no és necessària perquè és molt senzill calcular l'espai mostral E_B. Tanmateix, en la gran majoria dels casos conèixer E_B és realment complicat.

Exemple 23

Es llança el dau i el resultat ha estat un nombre parell. Troba la probabilitat que isca un 2 i que isca un 3.

Siga A = {obtenir un 2 en llançar un dau}
Siga B = {obtenir un nombre parell en llançar un dau} $E_B = \{2, 4, 6\}$
Siga C = {obtenir un 3 en llançar un dau}

Utilitzant E_B i no la fórmula:

$$\mathrm{P}(A/B) = \frac{\text{nre. casos favorables } A \cap B}{\text{nre. casos favorables en } B} = \frac{1}{3}$$

$$P(C/B) = \frac{\text{nre. casos favorables } A \cap B}{\text{nre. casos favorables en } B} = \frac{0}{3} = 0.$$

Utilitzant la fórmula:

És evident que $P(B) = \frac{3}{6} = \frac{1}{2}$. D'altra banda,

$A \cap B$ = {obtenir un 2} $\cap$ {obtenir un nombre parell} = {obtenir un 2},

$A \cap C$ = {obtenir un 3} $\cap$ {obtenir un nombre parell} = {Ø},

per la qual cosa $P(A \cap B) = \frac{1}{6}$ i $P(A \cap C) = 0$.

- $P(C / B) = \frac{P(C \cap B)}{P(B)} = \frac{0}{\frac{3}{6}} = 0$

- $P(A / B) = \frac{P(C \cap B)}{P(B)} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$

Probabilitat composta (teorema del producte)

Del concepte *probabilitat condicional* es dedueixen algunes conseqüències. La primera és el teorema del producte. El següent exemple l'introdueix.

Exemple 24

En una empresa hi ha 25 persones, de les quals hi ha 10 que saben més d'un idioma i la resta únicament en saben un. S'escull a l'atzar una persona, quina és la probabilitat que sàpia més d'un idioma?

La resposta a aquesta pregunta és relativament senzilla. Com que hi ha 10 persones que saben més d'un idioma i en total a l'empresa hi ha 25 treballadors:

$$P(\text{més d'un idioma}) = \frac{10}{25} = 0{,}4.$$

Però, si s'han d'escollir dues persones en lloc d'una, i s'escullen una a una, quina serà la probabilitat que totes dues sàpien més d'un idioma?

La resposta a aquesta pregunta és relativament més complexa. Així, el que demana el problema és:

P (que el primer sàpia més d'un idioma i el segon sàpia més d'un idioma).

Per a respondre aquesta segona qüestió cal tenir en compte els continguts que es donen a continuació:

De la definició de *probabilitat condicionada* es pot deduir que si E és un espai mostral, donats dos esdeveniments A i B de E de manera que $P(A) > 0$ i $P(B) > 0$, es compleix que:

$$P(A \cap B) = P(A / B) \cdot P(B)$$
$$P(B \cap A) = P(B / A) \cdot P(A)$$

Açò és així perquè per la definició de *probabilitat condicionada*,

$$P(A / B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(B \cap A) = P(A / B) \cdot P(B).$$

Anàlogament, $P(B / A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(B \cap A) = P(B / A) \cdot P(A)$.

Ara ja és possible contestar a la pregunta que encapçalava aquest epígraf:

P (que el primer sàpia més d'un idioma i el segon sàpia més d'un idioma) =

P (que el segon sàpia més d'un idioma/el primer sàpia més d'un idioma) · P (que el primer sàpia més d'un idioma) = $\frac{9}{24} \cdot \frac{10}{25} = \frac{90}{600} = \frac{1}{25} = 0{,}04$.

Si en compte de dos esdeveniments es tenen n esdeveniments $A_1, A_2, A_3, A_4, ..., A_n$ pertanyents a E:

$$P[\cap A_i] = P(A_1) \cdot P(A_2/A_1) \cdot P(A_3/A_1 \cap A_2) \cdot P(A_4/A_1 \cap A_2 \cap A_3) \cdot ... \cdot P(A_n/ \cap A_i).$$

Exemple 25

S'extrauen quatre boles i no són reemplaçades per altres d'una urna que en conté vuit de roges i deu de blaves. Calcula la probabilitat d'obtenir «blau, roig, roig, blau».

$$P(A_1 \cap R_2 \cap R_3 \cap A_4) = P(A_1) \cdot P(R_2/A_1) \cdot P(R_3/A_1 \cap R_2) \cdot P(A_4/A_1 \cap R_2 \cap R_3) =$$

$$= \frac{10}{18} \cdot \frac{8}{17} \cdot \frac{7}{16} \cdot \frac{9}{15} = 0{,}0686.$$

És interessant comentar que per a respondre a aquest tipus de preguntes es poden emprar arbres de camp, que en moltes ocasions simplifiquen i faciliten l'obtenció de les respostes. L'exemple següent ho posa de manifest.

Exemple 26

Una urna conté 10 boles verdes, 5 boles roges i 8 boles negres. S'extrauen tres boles aleatòriament de l'urna. Calcula la probabilitat d'obtenir la seqüència «verda, roja, negra» en els casos següents:

a) La bola extreta no es retorna a l'urna.
b) La bola extreta es retorna a l'urna.

Per a l'apartat *a*)

Definim: V_1 = la primera bola extreta és verda; R_2 = la segona bola extreta és roja; N_3 = la tercera bola extreta és negra.

El problema demana $P(V_1 \cap R_2 \cap N_3)$.

Aplicant-hi la definició de *probabilitat condicionada*, es calcula aquesta probabilitat:

$$P(V_1 \cap R_2 \cap N_3) = P(N_3/ V_1 \cap R_2) \cdot P(V_1 \cap R_2) =$$

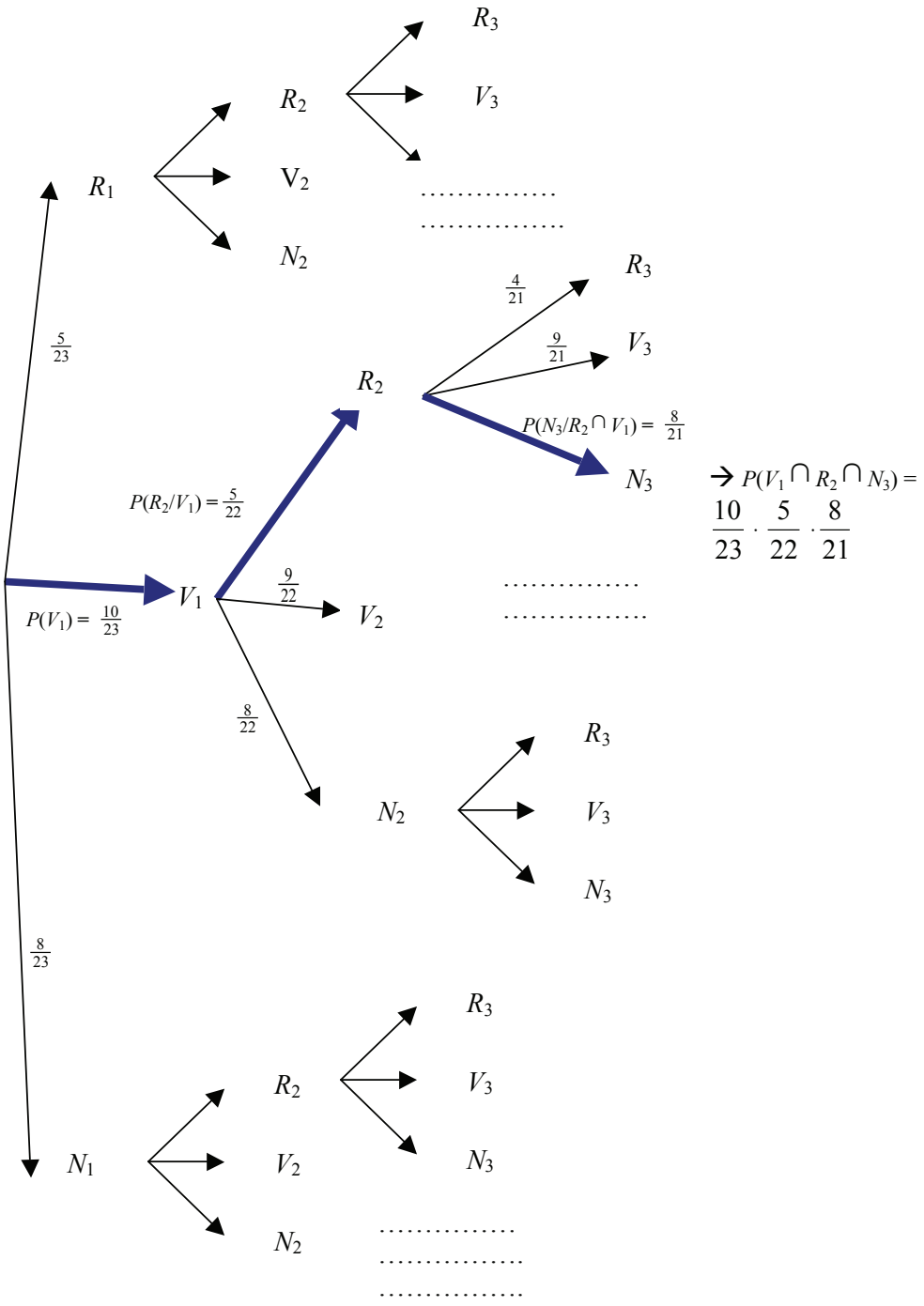
$$= P(N_3/ V_1 \cap R_2) \cdot P(R_2/V_1) \cdot P(V_1) = \frac{10}{23} \cdot \frac{5}{22} \cdot \frac{8}{21} = 0{,}0376,$$

Atès que:

$P(V_1) = \frac{10}{23}$ ja que a l'urna hi ha 23 boles, de les quals 10 són verdes.
$P(R_2/V_1) = \frac{5}{22}$ ja que a l'urna queden 22 boles –se n'ha extret una de verda– i en queden cinc de roges.

$P(N_3/V_1 \cap R_2) = \frac{8}{21}$ ja que a l'urna queden 21 boles –se n'han extret una de verda i una de roja– i en queden vuit de negres.

Aquest exercici es podria fer també en un arbre de camp com el següent:



El diagrama mostra el camí que permet obtenir la probabilitat demanada emprant l'arbre de camp.

Per a l'apartat *b*)

L'única diferència d'aquest exercici respecte de l'anterior és que la bola extreta es retorna a l'urna. Aquest fet implica que cada vegada que s'extrau una bola la composició de l'urna no varia. Per tant:

$P(V_1 \cap R_2 \cap N_3) = P(N_3/ V_1 \cap R_2) \cdot P(V_1 \cap R_2) = P(N_3/V_1 \cap R_2) \cdot P(R_2/V_1) \cdot \mathrm{P}(V_1)$ i com:

$P(V_1) = \frac{10}{23}$ ja que a l'urna hi ha 23 boles, de les quals 10 són verdes.

$P(R_2/V_1) = \frac{5}{23}$ ja que l'urna no varia. Que és el mateix que $P(R_2)$.

$P(N_3/ V_1 \cap R_2) = \frac{8}{23}$ ja que l'urna no varia. Que és el mateix que $P(N_3) =$

$= \frac{10}{23} \cdot \frac{5}{23} \cdot \frac{8}{23} = 0{,}0328.$

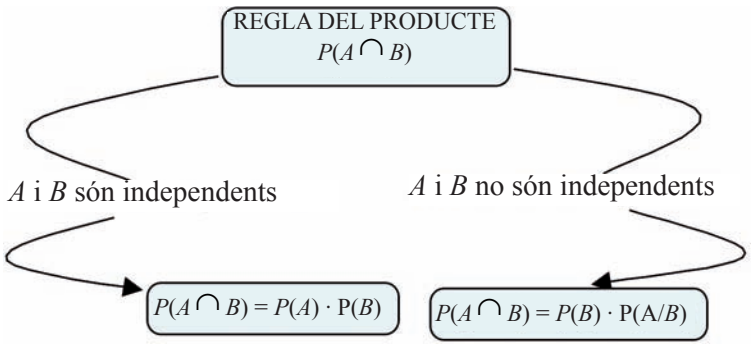
D'altra banda, s'observa que en aquests casos les condicions no afecten, és a dir, $P(R_2/V_1) = P(R_2)$; $\mathrm{P}(N_3/V_1 \cap R_2) = P(N_3)$. Llavors $P(V_1 \cap R_2 \cap N_3) = P(V_1) \cdot P(R_2) \cdot P(N_3)$.

L'arbre de camp que eixiria és semblant a l'anterior, però canviarien les probabilitats.

Nota

Aquest darrer exemple mostra una característica molt important de la teoria de la probabilitat condicionada. En el cas *b*), com que les boles són retornades a l'urna, l'elecció d'una bola no depèn ni té res a veure amb el resultat obtingut en les extraccions anterior ni posterior. Es podria dir que les extraccions de les boles són totes independents entre si. Aquest fet no ocorre en el primer cas de l'exemple, ja que cada extracció modifica el contingut de l'urna.

Així doncs, segons el que mostra l'exemple, sembla raonable anomenar *esdeveniments independents* aquests tipus d'esdeveniments en què l'ocurrència dels uns no influeix en l'ocurrència dels altres (apartat *b* de l'exemple anterior). A més a més, sembla clar que si dos esdeveniments són independents, aleshores la probabilitat de la intersecció d'ambdós esdeveniments és el producte de les probabilitats. El gràfic següent resumeix el que s'acaba de comentar. Posteriorment es posarà més èmfasi en aquestes qüestions.



6.5.2. Teorema de la probabilitat total

Una altra conseqüència de la probabilitat condicionada és el teorema que seguidament es presenta, i que permet calcular la probabilitat de l'un esdeveniment si es coneix una partició i les seues probabilitats.

Donat un espai mostral E, i un conjunt d'esdeveniments $A_1, A_2, ..., A_n \in \wp(E)$ *i partició de E, és a dir:*

- $\bigcup_{i=1}^{n} A_i = E$
- $A_i \cap A_j = \varnothing \quad \forall\, i \neq j$

Si es coneixen les probabilitats condicionades $P(B/A_i)$, $\forall$ i

$$\text{es té que } P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B/A_i) \cdot P(A_i)$$

Per a comprovar-ho, es considera que la figura 3 és l'espai mostral *E*.

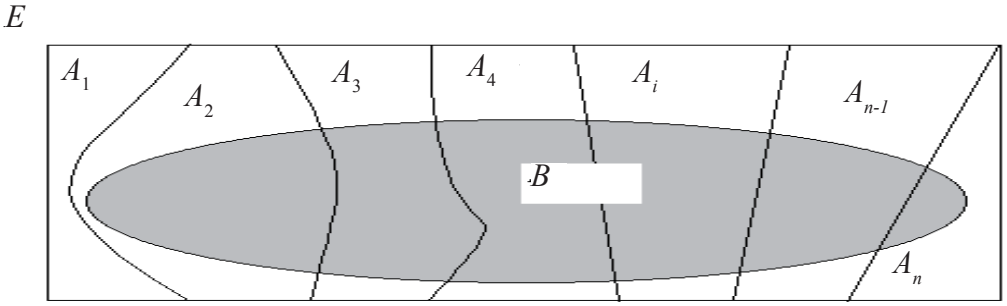


Figura 3

B es pot expressar: $B = (B \cap A_1) \cup (B \cap A_2) \cup (B \cap A_3) \cup ... \cup (B \cap A_i) \cup ... \cup (B \cap A_n)$

Com que tots els A_i són disjunts → tots els $(B \cap A_i)$ són també disjunts, llavors:

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) + ... + P(B \cap A_n).$$

I aplicant-hi el teorema del producte:

$$P(B) = P(B/A_1) \cdot P(A_1) + P(B/A_2) \cdot P(A_2) + ... + P(B/A_n) \cdot P(A_n)$$

$$= \sum_{i=1}^{n} P(B/A_i) \cdot P(A_i).$$

Exemple 27

Dues caixes contenen forrellats grans i xicotets. Se suposa que una caixa en conté 30 de grans i 10 de xicotets, i que l'altra en conté 30 de grans i 20 de xicotets. Se selecciona una caixa a l'atzar i s'extrau un forrellat. Quina és la probabilitat que el forrellat siga xicotet?

Siguen:

A_1 = seleccionar caixa 1
A_2 = seleccionar caixa 2
B = seleccionar forrellat xicotet

$P(B) = P(A_1) \cdot P(B/A_1) + P(A_2) \cdot P(B/A_2) = 1/2 \cdot 10/40 + 1/2 \cdot 20/50 = 0{,}125 + 0{,}2 = 0{,}325$.

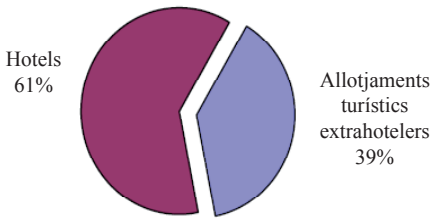
Exemple 28

Del total de les pernoctacions realitzades en allotjaments turístics col·lectius al mes d'agost a la Comunitat Valenciana, un 61% correspon a hotels i un 39% a allotjaments col·lectius extrahotelers. D'aquests darrers, la modalitat d'apartaments és la que presenta el nombre més gran de pernoctacions (50% del total), el 43% s'allotjaren en càmpings turístics i la resta en allotjaments de turisme rural.

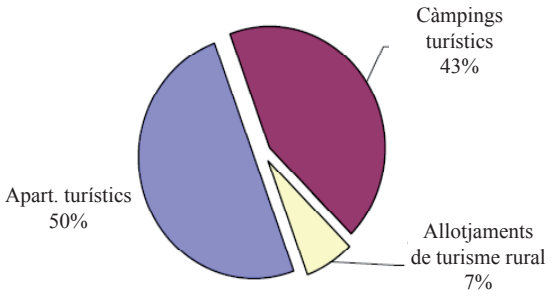
Si s'escull a l'atzar un turista que va pernoctar en durant el mes d'agost a la Comunitat Valenciana, quina és la probabilitat que estiguera en un allotjament de turisme rural? I en un càmping turístic?

Nota de premsa de l'INE. 29 d'agost de 2006.

Distribució de les pernoctacions d'allotjaments turístics en %



Distribució de les perncotacions d'allotjaments turístics extrahotelers per modalitat %



Una turista que s'allotjà a la Comunitat Valenciana durant el mes d'agost tenia dues possibilitats: allotjar-se en un hotel, amb probabilitat 0,61 o fer-ho en un allotjament turístic extrahoteler, amb una probabilitat del 0,39.

Aquests dos esdeveniments formen la partició de l'espai mostral perquè, bé es donà el primer cas, bé es donà el segon. A més a més, és impossible que ocorregueren les

dues coses al mateix temps (o s'allotjà en un hotel o no s'allotjà en un hotel). En conseqüència, és possible aplicar-hi el teorema de la probabilitat total.

Les probabilitats que es coneixen són:

- P(allotjament hotel) = 0,61
- P(allotjament extrahotel) = 0,39
- P(apart. turístic/allotjament extrahotel) = 0,5
- P(càmpings/allotjament extrahotel) = 0,43
- P(turisme rural/allotjament extrahotel) = 0,07
- P(apart. turístic/allotjament hotel) = 0
- P(càmpings/allotjament hotel) = 0
- P(turisme rural/allotjament hotel) = 0

Per tant:

P(turisme rural) = P(turisme rural/allotjament hotel) · P(allotjament hotel) +

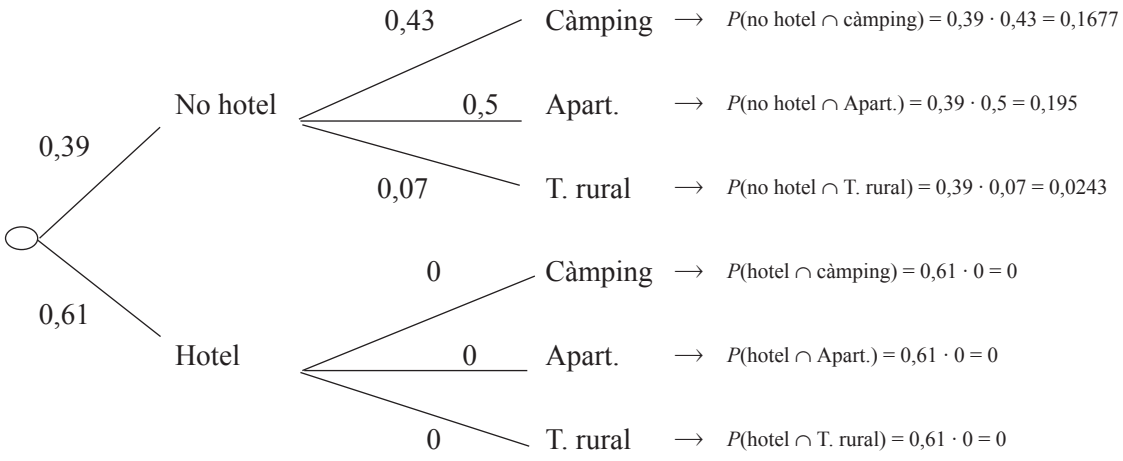
+ P(turisme rural/allotjament extrahotel) · P(allotjament extrahotel) =
= 0 · 0,61 + 0,07 · 0,39 = 0,0273.

Anàlogament:

P(càmping) = P(càmping/allotjament hotel) · P(allotjament hotel) +

+ P(càmping/allotjament extrahotel) · P(allotjament extrahotel) =
= 0 · 0,61 + 0,43 · 0,39 = 0,1677.

Cal dir que una manera més clara de fer l'estudi d'aquest tipus de problemes és mitjançant un arbre de camp que resumisca la informació:



Així doncs,

P(turisme rural) = P(turisme rural/allotjament hotel) · P(allotjament hotel)

+ P(turisme rural/allotjament extrahotel) · P(allotjament extrahotel)

= P(no hotel ∩ T. rural) + P(hotel ∩ T. rural) = 0,39 · 0,07 + 0,61 · 0 = 0,0243.

De la mateixa manera:

P(càmping) = P(càmping/allotjament hotel) · P(allotjament hotel)

+ P(càmping/allotjament extrahotel) · P(allotjament extrahotel)

= P(no hotel ∩ càmping) + P(hotel ∩ càmping) = 0 · 0,61 + 0,43 · 0,39 = 0,1677.

6.5.3. Teorema de Bayes

Una conseqüència del teorema anterior és el teorema de Bayes, que permet calcular probabilitats condicionades.

Siguen E un espai mostral i $\{A_i\} \in \wp(E)$ un conjunt d'esdeveniments de manera que:

- $\bigcup_i A_i \ \{A_i\} = E$,
- $A_i \cap A_j = \varnothing \ \ \forall \ i \neq j$,
- Són coneguts $P(A_i)$ per a qualsevol i, $P(A_i) > 0$
- Siga B un esdeveniment de manera que $P(B) > 0$ i del qual es coneixen $P(B/A_i)$ $\forall i$,

Llavors:

$$P(A_i/B) = \frac{P(B/A_i)\cdot P(A_i)}{P(B/A_i)\cdot P(A_1)) + P(B/A_2)\cdot P(A_2) + \ldots + P(B/A_n)\cdot P(A_n)}$$

És a dir:

$$P(A_i/B) = \frac{P(B/A_i)\cdot P(A_i)}{\sum_{i=1}^{n} P(B/A_j)\cdot P(A_j)}$$

Per a comprovar-ho, cal emprar el teorema anterior i la definició de *probabilitat condicionada*.

$$P(A_i/B) = \frac{P(A_i \cap B)}{P(A)}.$$

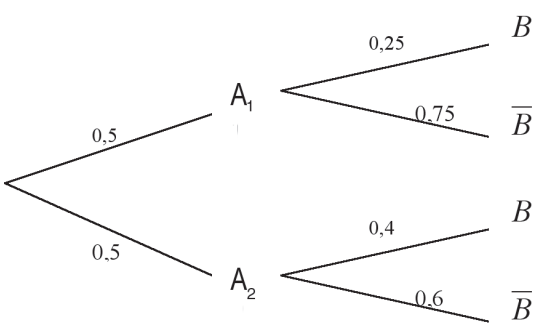
Si en el numerador s'aplica el teorema del producte, i en el denominador la probabilitat total, queda:

$$P(A_i/B) = \frac{P(A_i \cap B)}{P(A)} = \frac{P(B/A_i)\cdot P(A_i)}{\sum_{i=1}^{n} P(B/A_j)\cdot P(A_j)}.$$

Exemple 29

Considerant l'enunciat de l'exemple 27, així com tot el que hem desenvolupat, si el que ha ocorregut és que s'ha escollit un forrellat xicotet, quina és la probabilitat que aquest pertanya a la caixa 1?

L'arbre de camp següent resumeix la informació més rellevant de l'enunciat de l'exemple 27. És clar que el que es demana és $P(A_1/B)$. Aplicant-hi la definició de probabilitat condiconada:



$$P(A_1/B) = \frac{P(A1 \cap B)}{P(B)} =$$

$$= \frac{P(B/A1)\cdot P(A1)}{P(B/A2)\cdot P(A2) + P(B/A1)\cdot P(A1)} =$$

$$= \frac{0{,}125}{0{,}325} = 0{,}36$$

Exemple 30

Per a la fabricació d'un gran lot d'articles semblants s'utilitzen tres màquines: M_1, M_2 i M_3. La màquina 1 fabrica el 20% del articles, la màquina 2 el 30%, i la màquina 3 el 50% restant. La màquina 1 produeix un 1% d'articles defectuosos, la màquina 2 un 2%; i la màquina 3 un 3%. Se selecciona un article a l'atzar i resulta ser defectuós. Calcula la probabilitat que haja estat produït per la màquina 3.

Siguen:

D = ser defectuós
M_i = ser fabricat per M_i.

Així:

$P(M_1) = 0{,}2 \qquad P(M_2) = 0{,}3 \qquad P(M_3) = 0{,}5$
$P(D/M_1) = 0{,}01 \qquad P(D/M_2) = 0{,}02 \qquad P(D/M_3) = 0{,}03.$

Es demana la probabilitat de l'esdeveniment M_3/D. Es compleix que M_1, M_2 i M_3 formen una partició, per la qual cosa:

$$P(M_3/D) = \frac{P(D/M3)\cdot P(D)}{\sum_{i=1}^{3} P(D/M_i)\cdot P(M_i)}$$

$$= \frac{0{,}03 \cdot 0{,}5}{0{,}01 \cdot 0{,}2 + 0{,}02 \cdot 0{,}3 + 0{,}03 \cdot 0{,}5} = \frac{0{,}015}{0{,}023} = 0{,}06522.$$

6.5.4. Independència d'esdeveniments

En la nota de l'apartat 6.5.1 s'ha introduït el concepte *esdeveniments independents*. En aquest epígraf es formalitzarà el concepte i es donaran un conjunt de propietats, de les quals sens dubte, la més interessant és el teorema de caracterització. Aquest teorema confirma que les conjectures esmentades en la nota anterior són completament certes.

Definició

Es diu que dos esdeveniments A i B són estocàsticament independents si $P(A/B) = P(A)$, és a dir, que el fet que ocórrega l'esdeveniment B no influeix per a res en l'ocurrència de l'esdeveniment A.

El teorema següent caracteritza aquest tipus d'esdeveniments. És a dir, si dos esdeveniments el compleixen seran independents, i si són independents, llavors han de complir el teorema.

Teorema de caracterització

Dos esdeveniments A i B són independents –si i només si– $P(A \cap B) = P(A) \cdot P(B)$.

Demostració:
$(\Rightarrow)$

Se sap que $P(A \cap B) = P(A/B) \cdot P(B)$. Com que se suposa que A i B són independents, es té que $P(A/B) = \mathrm{P}(B)$.

Unint ambdues coses, $P(A \cap B) = P(A) \cdot P(B)$.

($\Leftarrow$)

Ara se suposa que $P(A \cap B) = P(A) \cdot P(B)$. Com que $P(A \cap B) = P(A/B) \cdot P(B)$, substituint:

$P(A) \cdot P(B) = P(A/B) \cdot P(B)$. Per tant, $P(A) = P(A/B)$, i així els esdeveniments A i B són independents.

Exemple 31

Es considera el senzill experiment que consisteix a llançar dues vegades un dau.

Si definim els experiments:

A = el resultat del primer dau mostra una xifra senar
B = el resultat del segon dau mostra una xifra menor que 3.

Són independents?

Emprant la definició

És evident que, intuïtivament, els esdeveniments A i B semblen independents, l'ocurrència de l'un no afecta, en absolut, en l'ocurrència de l'altre. Però, comprovem-ho:

L'espai mostral de l'experiment està format per 36 casos possibles: E ={(1, 1), (1, 2), ..., (1, 6), (2, 1), (2, 2), (2, 3), ..., (2, 6), ..., (6, 1), (6, 2), ..., (6, 6)}. D'aquests, els esdeveniments A i B estan formats per:

A = {(1, 1), (1, 3), (1, 5), (2, 1), (2, 3), (2, 5), ..., (6, 1), (6, 3), (6, 3)}; n'hi ha 18 casos.

B = {(1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (1, 2), (2, 2), (3, 2), ..., (6, 2)}; n'hi ha 12 casos.

I la intersecció:

$A \cap B$ = {(1, 1) , ((3, 1) , (5, 1) , (1, 2) , (3, 2) , (5, 2)}; n'hi ha 6 casos.

Llavors, $P(A) = \frac{18}{36} = \frac{1}{2}$, $P(B) = \frac{12}{36} = \frac{1}{3}$ i $P(A \cap B) = \frac{6}{36} = \frac{1}{6}$.

Ara bé, tenint en compte la definició de *probabilitat condicionada*:

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{3}} = \frac{1}{2} = P(A).$$

Així doncs, la probabilitat de A condicionada a B, com s'havia predit, és el mateix que la $P(A)$.

Anàlogament:

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3} = P(B)$$. I, per tant, $P(B/A)$ és igual a $P(B)$.

Emprant el teorema de caracterització

El fet que aquests dos esdeveniments són independents també s'hauria pogut comprovar usant el teorema de caracterització, ja que:

$P(A \cap B) = \frac{1}{6}$ i $P(A) \cdot P(B) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$ $P(A \cap B) = P(A) \cdot P(B)$ i, en virtut del teorema de caracterització, A i B són independents.

Veurem tot seguit un exemple de dos esdeveniments que no són independents.

Exemple 32

Es considera l'experiment «llançar una moneda dues vegades». Es consideren els esdeveniments I = «el primer llançament ha estat cara» i J = «els dos llançaments han estat cara». Són I i J esdeveniments independents?

Per a comprovar-ho, emprarem el teorema de caracterització. Així:

L'espai mostral $E = \{(C, C); (C, +); (+, C); (+, +)\}$

L'esdeveniment $I = \{(C, C); (C, +)\}$; l'esdeveniment $J = \{(C, C)\}$ i $I \cap J = \{(C,C)\}$.

A més a més: $P(I \cap J) = \frac{1}{4}$; $P(I) = \frac{1}{2}$; i $P(J) = \frac{1}{4}$ com que $P(I \cap J) \neq \frac{1}{8} = P(I) \cdot P(J)$, els esdeveniments I i J no són independents.

Per altra part, emprant el teorema de caracterització es pot estendre el concepte *independència* a tres esdeveniments. Així, es diu que tres esdeveniments A_1, A_2 i A_3 són independents si –i només si– verifiquen les relacions:

- $P(A_1 \cap A_2) = P(A_1) \cdot P(A_2)$
- $P(A_1 \cap A_3) = P(A_1) \cdot P(A_3)$
- $P(A_2 \cap A_3) = P(A_2) \cdot P(A_3)$ i
- $P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3)$

Aquesta definició es pot generalitzar a n esdeveniments. Un conjunt d'esdeveniments $\{A_1, A_2, ..., A_n\}$ són independents si per a qualsevol subconjunt d'aquest $\{A_i, A_j, ..., A_m\}$ es compleix que $P(A_i \cap A_j \cap ... A_m) = P(A_i) \cdot P(A_j) \cdot ... \cdot P(A_m)$.

Nota

Una de les qüestions que més sol confondre els estudiants d'estadística pel que fa a la independència d'esdeveniments és la següent: són els esdeveniments disjunts independents?

La resposta és clara: no. Dos esdeveniments disjunts són aquells que tenen intersecció buida. És a dir, si l'un ocorre l'altre no pot fer-ho (l'un depèn de l'altre). Formalment també és evident. Si A i B són dos esdeveniments disjunts, tots dos diferents de $\varnothing$, llavors $P(A) > 0$ i $P(B) > 0$, $P(A) \cdot P(B) > 0$. Tanmateix, $P(A \cap B) = P(\varnothing) = 0$; pel teorema de caracterització, A i B no són independents.

Nota

Hi ha experiments que estan formats per la repetició d'experiments de caràcter més simple –com per exemple llançar dues vegades una moneda– i sobre els que es pot fer l'anàlisi de dues maneres: d'una banda, es pot considerar com un únic experiment i d'una altra, com una mena de concatenació dels experiments simples que el formen.

Així, un exemple molt simple és l'esmentat anteriorment, que consiteix a llançar dues vegades una moneda equilibrada. L'espai mostral és $E = \{(C, C); (C, +); (+, C); (+, +)\}$ i si l'analitzem com un únic experiment, obtenim que tots els esdeveniments elementals són equiprobables per ser la moneda equilibrada. Llavors:

$$P(C, C) = P(C, +) = P(+, C) = P(+, +) = \frac{1}{4}.$$

Fent l'anàlisi com si foren dos experiments, en cadascun $P(C) = P(+) = \frac{1}{2}$ i l'espai mostral es pot calcular combinant cada resultat del primer experiment simple amb cada resultat del segon: $E^1 = \{(C_1, C_2); (C_1,+2); (+1, C_2); (+1, +2)\}$. A més a més, les probabilitats dels elements de l'espai mostral es poden calcular emprant la regla del producte de probabilitats:

$P(C, C) = P(C_1) \cdot P(C_2/C_1) = P(C_1) \cdot P(C_2)\ \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$; per independència de C_2 i C_1.

$P(+, C) = P(+_1) \cdot P(C_2/+_1) = P(+_1) \cdot P(C_2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$; per independència de C_2 i $+_1$.

1. Els subíndexs indiquen l'ordre temporal en què ocorren els esdeveniments. A més a més, cal notar que el parell ordenat (A, B) representa A∩B, per això cal emprar els subíndexs.

$P(C,+) = P(C_1) \cdot P(+_2/C_1) = P(C_1) \cdot P(+_2) \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$; per independència de $+_2$ i C_1.

$P(+,+) = P(+_1) \cdot P(+_2/+_1) = P(+_1) \cdot P(+_2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. per independència de $+_2$ i $+_1$.

Fent l'anàlisi d'aquesta manera s'observa que, més que de *esdeveniments independents*, es pot parlar de *experiments independents*. I si això és així, els resultats d'un són independents dels resultats dels altres. A més a més, emprant aquesta òptica el càlcul de les probabilitats és més senzill.

Per exemple, si considerem l'experiment que consisteix a llançar tres vegades una moneda equilibrada i calcular la probabilitat d'obtenir (*C*, *C*, +), l'anàlisi canvia si es realitza d'una manera o d'una altra:

Forma 1

Cal calcular l'espai mostral. Si no és possible, almenys cal saber el nombre d'esdeveniments elementals que el formen per a poder aplicar Laplace. En aquest cas, l'espai mostral:

$E = \{(C, C, C); (C, C, +); (C, +, C); (C, +, +); (+, +, +); (+, +, C); (+, C, +); (+, C, C)\}$

$$P(C, C, +) = \frac{\text{casos favorables}}{\text{casos possibles}} = \frac{1}{8}$$

Forma 2

Com que els tres experiments simples són independents, llavors:

$$P(C, \mathrm{C}, +) = \mathrm{P}(C_1) \cdot P(C_2) \cdot P(+_3) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}.$$

Si en lloc de considerar tres llançaments se'n consideren trenta, la dificultat per a calcular l'espai mostral augmentaria i caldria recórrer a la combinatòria per a esbrinar el nombre de casos possibles i casos favorables. Seria, doncs, més recomanable considerar l'experiment com una concatenació de trenta experiments independents simples on en cadascun la probabilitat d'obtenir cara en cada llançament és constant i no depèn del resultat obtingut en el llançament anterior.

La pregunta que ara ens podem fer és: aquest mètode d'anàlisi únicament es pot realitzar en experiments independents? L'exemple següent respon a aquesta qüestió.

Exemple 33

En un grup de 15 amics n'hi ha 8 que treballen en un empresa del sector del taulell, 4 que treballen en l'agricultura i la resta estan actualment a l'atur. Si s'escullen a l'atzar 2 persones del grup, quina és la probabilitat que la primera escollida estiga a l'atur i la segona treballe en el sector del taulell?

En primer lloc, pot considerar-se l'experiment «extraure dues persones del grup» com un experiment compost? La resposta és que sí, ja que en cada experiment simple els possibles resultats són {aturat (*A*), sector del taulell (*ST*), agricultura(*AG*)} i l'espai mostral de l'experiment és $E = \{(A, A); (A, ST); (A, G); (ST, A); (ST, AG); (ST, ST); (AG, A); (AG, ST); (AG, AG)\}$ es calcula combinant els espais mostrals dels experiments simples.

A més a més, $P(A, ST) = P(A_1, ST_2) = P(ST_2/A_1) \cdot P(A_1) = \frac{8}{14} \cdot \frac{3}{15}$.

Així doncs, no és que l'anàlisi dels experiments compostos des del punt de vista dels simples que els conformen es puga fer encara que no hi haja independència, sinó que és del tot recomanable en aquests casos. La raó és que l'espai mostral de l'experiment compost no és equiprobable, i aquest fet complica molt el càlcul de les probabilitats d'esdeveniments elementals.

Una altra qüestió que ens podem preguntar és: en qualsevol experiment s'ha de comprovar sempre la independència dels esdeveniments? La resposta és que en molts casos es pot suposar que són independents si existeixen raons de pes que així ho indiquen. Per exemple, en el llançament d'un dau dues vegades, com que l'experiment es repeteix en les mateixes condicions inicials, el resultat de la primera tirada es pot suposar que és independent del resultat del segon llançament.

En general, les condicions en què es realitzen els experiments, així com les seues característiques, fan possible determinar si la suposició d'independència es justifica o si almenys ho és aproximadament (exemple 34).

Exemple 34

En una determinada regió, 8.000 persones treballen en el sector del taulell, 4.000 treballen en l'agricultura i 2.000 estan actualment a l'atur. Si s'escullen a l'atzar 2 persones del grup, quina és la probabilitat que la primera escollida estiga a l'atur i la segona treballe en el sector del taulell?

Aquest problema és exactament el mateix que l'anterior, l'únic que varia és el nombre de persones. Per tant, és un experiment compost en què no es dóna independència dels simples. Llavors:

$$P(A, ST) = P(A_1, ST_2) = \mathrm{P}(ST_2/A_1) \cdot P(A_1) = \frac{8000}{13.999} \cdot \frac{2000}{14.000} = 0{,}08163848.$$

Però per altra part, en aquest cas podem considerar que els experiments simples són independents, ja que el fet de traure una persona de l'espai mostral, com que aquest és molt gran, afecta reltivament poc el càlcul de probabilitats. Així, assumint independència:

$$P(A, ST) = P(A_1, ST_2) = P(A_1) \cdot P(ST_2) = \frac{8000}{14.000} \cdot \frac{2000}{14.000} = 0{,}081632653,$$

s'obté un error absolut inferior a 0,00001 i es pot considerar menyspreable.

Així doncs, en aquest exemple la suposició d'independència no altera en excés els resultats i, per tant, el benefici obtingut en simplicitat del problema i l'escàs error, recomanen en aquests casos considerar els esdeveniments independents.

Nota

Redundància?

Un article sobre els avions antiradar aparegut a la revista *Popular Science* afirma: «Un avió construït en bona part amb material de fibra de carboni fou el Lean Fan 2100, que havia de dur dos transponedors de radar. La raó és que si fallava un sol transponedor l'avió era casi invisible per al radar». Aquesta redundància és una aplicació de la teoria del producte de la probabilitat, ja que si un component té una probabilitat de fallar de 0,001, la probabilitat que fallen dos components és de 0,0000001.

Per a finalitzar aquest epígraf s'enunciaran, sense demostrar, un seguit de propietats de la independència d'esdeveniments que poden ser d'utilitat a l'hora d'analitzar problemes.

Propietats de la independència estocàstica:

Si A i B són independents $\Rightarrow$ $\overline{A}$ i B també ho són.

Si A i B són independents $\Rightarrow$ A i $\overline{B}$ també ho són.

Si A i B són independents $\Rightarrow$ $\overline{A}$ i $\overline{B}$ també ho són.

Si hi ha implicació entre A i B $\Rightarrow$ No hi ha independència
(llevat que $A = E$ o $B = E$).

6.6. Problemes proposats

En aquest epígraf es plantejaran un conjunt de problemes per a la resolució dels quals és necessari conèixer la teoria desenvolupada al llarg de la unitat.

Exercici 1

Per a l'experiment «llançar dos daus consecutivament»:

a) Troba els diferents resultats de l'experiment (espai mostral).
b) Digues els esdeveniments elementals que formen l'esdeveniment compost: A = la suma de les xifres és parella. Calcula el seu complementari: $\overline{A}$.
c) Digues els esdeveniments elementals que formen l'esdeveniment compost: C = la suma de les xifres és superior a 5. Calcula el seu complementari $\overline{C}$.
d) Calcula l'esdeveniment $A - \mathrm{C}$.
e) Calcula l'esdeveniment $A \cup C$ i $A \cap C$.
f) Calcula l'esdeveniment $(A \cup \mathrm{C}) \cup (A \cap C)$.

Exercici 2

En una empresa de Castelló treballen 200 persones. 80 saben parlar anglès, 110 francès i 60 alemany. 50 persones saben parlar francès i anglès, 35 saben parlar francès i alemany, i 40 saben parlar anglès i alemany. A més a més, 30 treballadors saben parlar tots tres idiomes.

En aquestes condicions:

a) Calcula la probabilitat que si s'escull a l'atzar una persona de l'empresa, únicament sàpia parlar anglès.
b) Calcula la probabilitat que si s'escull a l'atzar una persona de l'empresa, no sàpia parlar cap dels tres idiomes.

Exercici 3

En una oposició els aspirants han de desenvolupar un tema. Han d'escollir-ne un dels cinc que el tribunal tria de manera aleatòria d'un conjunt de 72 temes. Si una persona es presenta a l'oposició i se n'ha estudiat 30:

a) Quina és la probabilitat que s'haja estudiat almenys un dels cinc temes que ha extret el tribunal? I que n'haja estudiat almenys dos dels cinc?
b) Quina és la probabilitat que no se n'haja estudiat cap dels cinc?

Exercici 4

En una empresa treballen 300 dones i 700 homes. De les entrevistes de treball se sap que 450 persones parlen anglès, 375 persones parlen francès i n'hi ha 500 que no saben parlar cap dels dos idomes. S'hi escull una persona a l'atzar.

a) Quina és la probabilitat que siga dona?
b) Quina és la probabilitat que sàpia anglès? I que sàpia francès?
c) Quina és la probabilitat que sàpia parlar almenys un dels dos idiomes?
d) Quina és la probabilitat que sàpia parlar tots dos idiomes?

Exercici 5

Una urna conté 10 boles verdes, 5 boles roges i 8 boles negres. S'extrauen 3 boles de l'urna aleatòriament. Calcula la probabilitat d'obtenir la seqüència «verda, roja, negra» en els casos següents:

a) La bola extreta no es retorna a l'urna.
b) La bola extreta es retorna a l'urna.

Exercici 6

La taula mostra l'alumnat estranger classificat per ensenyaments a Espanya al llarg del curs 2004/2005. Aquests ensenyaments es poden classificar en tres grups atenent l'edat de l'alumne i el grau de maduresa. Així, el grup 1 està format pels ensenyaments educació infantil, educació primària i educació especial. El grup 2, per educació secundària obligatòria i pel batxillerat. El grup 3, per la resta d'ensenyaments. Es considera l'experiment «seleccionar aleatòriament un alumne/a estranger d'Espanya l'any 2004/2005 i preguntar-li a quin grup pertany».

	2004-05
TOTAL	**459.291**
Educación Infantil/ Preescolar	85.799
Educación Primaria/ E.G.B	198.165
Educación Especial	1.572
E.S.O	124.714
Bachilleratos	19.160
Formación Profesional	19.255
Enseñanzas Artísticas	1.816
Enseñanzas de Idiomas	8.795
Enseñanzas Deportivas	15
No consta enseñanza (Enseñanzas Régimen General)	0

Fuente de información: Ministerio de Educación y Ciencia.

Font: Anuari 2007, INE

a) Quin és l'espai mostral d'aquest experiment? Els esdeveniments elementals tenen la mateixa probabilitat de donar-se?
b) Calcula la probabilitat de cada esdeveniment elemental.
c) Quina relació hi ha entre la probabilitat del'esdeveniment«cursar ESO» i l'esdeveniment «pertànyer al grup 2»?

Exercici 7

La taula següent mostra dades referents a la Comunitat Valenciana.

	Población a 1 de enero de 2006		
	Total	Mujeres (%)	Extranjeros (%)
Comunitat Valenciana	**4.806.908**		
Alicante/Alacant	1.783.555	49,9	20,1
Castellón/Castelló	559.761	49,7	13,9
Valencia/València	2.463.592	50,5	9,4

Font: INE

S'ha escollit a l'atzar una persona de la Comunitat i ha resultat ser home. Calcula la probabilitat que siga de Castelló.

TEMA 7

Introducció a la probabilitat (II): models de probabilitat discrets i continus

Objectius tema 7

- Reconèixer característiques comunes en els experiments aleatoris senzills.
- Conèixer els conceptes *variable aleatòria* i *model de probabilitat*.
- Conèixer els trets més característics de les distribucions de probabilitat: funcions de densitat o de probabilitat, funcions de distribució, esperança matemàtica, variància, etc.
- Identificar els experiments que es modelen mitjançant una distribució discreta.
- Conèixer i saber fer càlculs amb les distribucions discretes de Bernoulli, binomial, hipergeomètrica i de Poisson.
- Identificar els experiments que es modelen mitjançant una distribució contínua.
- Conèixer i saber fer càlculs amb les distribucions contínua, uniforme, exponencial i normal.
- Conèixer i saber aplicar els teoremes d'aproximacions de diferents distribucions per la distribució normal, fins i tot el teorema del límit central.

7.1. Introducció

En la unitat anterior s'han tractat diverses qüestions relatives a la probabilitat. Així, donat un experiment aleatori es determinaven les probabilitats de l'ocurrència de cadascun dels esdeveniments. Per a fer-ne els càlculs, calia conèixer les característiques bàsiques de l'experiment i aplicar-hi els axiomes i les propietats de la probabilitat.

El coneixement d'estudis probabilístics d'experiments semblants a un de donat, únicament permetia intuir els càlculs de les probabilitats de l'experiment en qüestió. Per exemple, els experiments «llançar una moneda equilibrada i observar-ne el resultat» i «extraure una bola d'una urna amb tres boles, dues de verdes i una de roja i observar el color de la bola extreta» són molt semblants, però, amb el que s'ha estudiat en el capítol anterior, tan sols es poden utilitzar els càlculs de probabilitats del primer com un exemple per a realitzar-ne els càlculs en el segon.

D'altra banda, és evident que tots dos experiments tenen característiques molt semblants des del punt de vista de la probabilitat:

- Són aleatoris.
- Tenen dos resultats possibles (C, +) i (verd, roig).
- Les probabilitats de cada esdeveniment elemental són P(esdeveniment 1) = p i P(esdeveniment 2) = $1 - p$.

$$P(C) = \frac{1}{2} \text{ i } P(+) = 1 - \frac{1}{2} = \frac{1}{2} \qquad \text{i} \qquad P(\text{verd}) = \frac{2}{3} \text{ i } P(\text{roig}) = 1 - \frac{2}{3} = \frac{1}{3}$$

Així, tots dos experiments segueixen un mateix patró en la distribució de probabilitats en relació als esdeveniments elementals –coneguda la probabilitat d'un esdeveniment elemental, la probabilitat de l'altre és 1 menys la probabilitat del primer–, és a dir, tots dos segueixen el mateix model de probabilitat. Cal dir que existeixen molts experiments que no segueixen aquest patró, per exemple, llançar un dau equilibrat i observar el resultat.

Al llarg del capítol s'estudiaran conceptes com ara *variable aleatòria*, *funció de probabilitat*, *funció de quantia*, *funció de distribució*, *distribució de probabilitat*, etc., els quals permeten assignar cada experiment a un model de probabilitat. També es presentaran els principals models de probabilitat: de Bernoulli, binomial, de Poisson, uniforme, exponencial i el model normal. Per a finalitzar s'introduirà el teorema del límit central com una de les aplicacions més importants de la distribució normal.

7.2. De l'experiment al model

Hi ha nombrosos experiments aleatoris que posseeixen característiques intrínseques comunes des del punt de vista de la probabilitat, malgrat que els seus resultats i conseqüències siguen diferents.

Així, en comparar els experiments:

> Exp. 1: llançar 15 vegades una moneda equilibrada i calcular la probabilitat d'obtenir 8 cares en els 15 llançaments.
>
> Exp. 2: triar a l'atzar 8 persones d'una mostra de 15 i calcular la probabilitat que totes siguen consumidores d'un determinat producte.

S'observa que en ambdós casos es realitzen 15 experiments consecutius i independents els uns dels altres. El resultat de cadascun d'aquests 15 experiments té dues úniques possibles opcions: en Exp. 1, cara o creu, i en Exp. 2, ser consumidor del producte o no ser consumidor del producte. A més, de la mateixa manera que la probabilitat d'obtenir cara en el primer experiment és 0,5 (la moneda és equilibrada), en el segon experiment és segur que ha d'existir una probabilitat **p** que una persona triada a l'atzar de la mostra siga consumidora del producte.

Per tant, es pot dir que ambdós experiments segueixen un mateix patró, ambdós poden modelar-se per la mateixa distribució de probabilitat.

Vegem el significat d'açò últim per mitjà d'un exemple:

Exemple 1

Experiment 3: llançar tres vegades una moneda equilibrada i comptar el nombre de cares que apareixen.

L'espai mostral de l'experiment «llançar una moneda tres vegades» és:

E: {(*C*, *C*, *C*), (*C*, *C*, +), (*C*, +, *C*), (+, *C*, *C*), (+, +, *C*), (+, *C*, +), (*C*, +, +), (+, +, +)}

Es representa per *X* allò que realment es vol estudiar, és a dir,

X = nombre de cares que apareixen en tots tres llançaments.

Quins valors pot prendre X? Únicament pot prendre els valors 0, 1, 2 i 3.

Així:

$X = 3$ si es dóna l'esdeveniment (C, C, C).
$X = 2$ si es dóna l'esdeveniment $((C, C, +)$ o $(C, +, C)$ o $(+, C, C))$.
$X = 1$ si es dóna l'esdeveniment $((+, C, +)$ o $(C, +, +)$ o $(+, +, C))$.
$X = 0$ si es dóna l'esdeveniment $(+, +, +)$.

Per tant, l'assignació de probabilitats serà:

$$P(X = 0) = P((+, +, +)) = \frac{1}{8}$$

$$P(X = 1) = P((+, +, C) \cup (+, C, +) \cup (C, +, +)) = \frac{3}{8}$$

$$P(X = 2) = P((C, C, +) \cup (C, +, \mathrm{C}) \cup (+, C, C)) = \frac{3}{8}$$

$$P(X = 3) = P((C, C, C)) = \frac{1}{8}.$$

Lògicament, es pot calcular la probabilitat de qualsevol esdeveniment a partir dels valors obtinguts:

Es considera $A = \{X$ estrictament inferior a $2\}$

$$P(X \in A) = P(X \in \{0, 1\}) = P(X = 0 \text{ o } X = 1)$$

$$= P(X = 0) + P(X = 1) = \frac{1}{8} + \frac{3}{8} = \frac{4}{8} = \frac{1}{2}.$$

Aquest experiment pot generar un model de distribució de probabilitats.

Així, qualsevol experiment en què:

- es realitzen tres experiments consecutius i independents els uns dels altres, amb dos possibles resultats equiprobables, i

- es desitge conèixer la probabilitat que apareguen 0, 1, 2 o 3 vegades un d'aquests resultats,

admet la distribució de probabilitats següent: si es denomina $X =$ nombre de vegades que es dóna el resultat desitjat en tots tres experiments:

$$P(X = 0) = \frac{1}{8} \quad P(X = 1) = \frac{3}{8}.$$

$$P(X = 2) = \frac{3}{8} \quad P(X = 3) = \frac{1}{8}.$$

Es pot utilitzar el model teòric en qualsevol experiment que complisca els requisits. Vegem-ne un exemple:

Exemple 2

Experiment 4:

Se sap que dues màquines A i B d'una empresa fabriquen bolígrafs iguals, que es guarden en caixes (se suposa que en cada caixa hi ha la mateixa quantitat de bolígrafs procedents de la màquina A que de la B).

Es trien a l'atzar tres bolígrafs de tres caixes diferents, i es desitja saber la probabilitat que almenys dos hagen sigut fabricats per la màquina B.

S'observa que aquest experiment compleix les dues premisses anteriors:

- Són tres experiments independents i consecutius, cada un amb dos resultats possibles i equiprobables:

$$P(\text{ser de la màquina A}) = P(\text{ser de la màquina B}).$$

- Es desitja conèixer la probabilitat que apareguen 0, 1, 2 o 3 bolígrafs d'una de les màquines.

Es denomina X = nombre de bolígrafs que procedeixen de la màquina B.

Es vol saber $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}$.

Aquest càlcul es pot realitzar d'una altra manera: $P(X \leq 2) = 1 - P(X > 2) = 1 - \frac{1}{8} = \frac{7}{8}$.

7.3. Variables aleatòries. Estudi de la seua distribució

El que es pretén amb l'ús de variables aleatòries és transformar l'estudi dels experiments per facilitar-ne la comprensió i la recerca del model de probabilitat que segueix. Així, bàsicament les variables aleatòries són funcions o relacions que assignen a cada esdeveniment elemental d'un experiment aleatori un nombre real. Aquest fet provoca una equivalència entre l'espai mostral i el conjunt numèric producte de les assignacions.

Així doncs, s'anomena *variable aleatòria* aquella funció que assigna a cada esdeveniment elemental de l'espai mostral un nombre. És a dir, si E és l'espai mostral, X és una variable aleatòria si és una funció que associa a cada esdeveniment elemental **s** de E un nombre $X(s)$. El conjunt format per tots els possibles valors que pren la variable aleatòria s'anomena *rang* (R) de la variable aleatòria.

D'altra banda, donada una variable aleatòria X i un subconjunt R_x del rang de X, la probabilitat de R_x està donada per la probabilitat que es done en el conjunt format pels esdeveniments elementals A, els quals tenen per imatge de X, R_x, és a dir, $X(A) = R_x$. Els exemples següents aclareixen aquestes qüestions.

Exemple 3

Es considera l'experiment «escollir una persona d'una ciutat en la qual viuen persones de quatre nacionalitats: França (25.000 persones), Espanya (500.000 persones), Itàlia (25.000 persones) i Grècia (25.000 persones) i observar de quin país és».

En aquest exemple l'espai mostral és {França, Espanya, Rússia i Grècia} i la distribució de probabilitats és:

$$P(\text{ser de França}) = \frac{25.000}{125.000} = \frac{1}{5} \qquad P(\text{ser d'Espanya}) = \frac{50.000}{125.000} = \frac{2}{5}$$

$$P(\text{ser de Rússia}) = \frac{25.000}{125.000} = \frac{1}{5} \qquad P(\text{ser de Grècia}) = \frac{25.000}{125.000} = \frac{1}{5}$$

Pot definir-se la variable aleatòria X, de manera que es produïsquen les assignacions: X(França) = 1, X(Espanya) = 0, X(Grècia) = 2 i X(Rússia) = 3.

Per tant, l'espai mostral es transforma en el conjunt numèric = {0, 1, 2, 3}.

També es calculen les probabilitats corresponents dels valors que pren la variable X:

$$P(X = 0) = P(\text{ser de Espanya}) = \frac{50.000}{125.000} = \frac{2}{5}$$
$$P(X = 1) = P(\text{ser de França}) = \frac{25.000}{125.000} = \frac{1}{5}$$
$$P(X = 2) = P(\text{ser de Grècia}) = \frac{25.000}{125.000} = \frac{1}{5}$$
$$P(X = 3) = P(\text{ser de Rússia}) = \frac{25.000}{125.000} = \frac{1}{5}.$$

D'aquesta manera s'obté l'experiment aleatori transformat.

Exemple 4

Es considera l'experiment «preguntar a una persona si li agrada un determinat article de consum» amb les possibles respostes «gens», «poc», «regular», «prou», «molt». Es pot definir la variable aleatòria X, de manera que:

$$X\,(\text{gens}) = 1,\ X\,(\text{poc}) = 2,\ X\,(\text{regular}) = 3,\ X\,(\text{prou}) = 4,\ X(\text{molt}) = 5.$$

Evidentment la distribució de probabilitats serà:

$P(X = 1) = P(\text{gens})$ $\quad$ $P(X = 2) = P(\text{poc})$
$P(X = 3) = P(\text{regular})$ $\quad$ $P(X = 4) = P(\text{prou})$
$P(X = 5) = P(\text{molt})$

Exemple 5

En l'experiment 3 d'aquest epígraf, també s'ha definit la variable aleatòria X = nombre de cares. En el mateix experiment s'ha calculat també la distribució de probabilitats.

Nota

Cal dir que en moltes ocasions els possibles resultats de l'experiment són numèrics i no cal, doncs, definir una variable aleatòria. Per exemple en l'experiment «observar el temps que tarda un alumne a realitzar un exercici d'estadística» l'espai mostral està format per nombres reals.

Exemple 6

Siga l'experiment «llançar dues vegades un dau en forma de tetràedre amb cares iguals a 1, 2, 3 i 4 i sumar els resultats».

Com que en aquest cas el resultat de l'experiment és un nombre, es considera que la variable aleatòria és el mateix resultat de l'experiment. Així la variable aleatòria pren els valors {2, 3, 4, 5, 6, 7, 8} i les probabilitats de cada valor són:

$$\left\{\frac{1}{16},\frac{1}{8},\frac{3}{16},\frac{1}{4},\frac{3}{16},\frac{1}{8},\frac{1}{16}\right\} \text{ respectivament.}$$

(La comprovació es deixa com a exercici.)

Tipus de variables aleatòries

De la mateixa manera que ocorria amb les variables estadístiques estudiades en el bloc referent a estadística descriptiva, les variables aleatòries poden ser discretes o contínues.

Les variables aleatòries discretes poden assumir tan sols certs valors, amb freqüència nombres enters, i resulten fonamentalment del recompte. La variable «resultat obtingut en llançar un dau» és discreta i els possibles valors que pot prendre són {1, 2, 3, 4, 5 i 6}. També el nombre de trucades que rep una família en una hora, el nombre de vaixells que arriben a un port o el nombre de clients que tenen les empreses ceràmiques són exemples de variables aleatòries discretes.

Les variables aleatòries contínues poden prendre qualsevol valor dins d'un rang donat. La variable *temps que tarda un alumne a realitzar un examen amb un límit de temps d'una hora* és contínua, i els valors que pot prendre són tots aquells compresos entre 0 i 60 minuts. Així 30,252 minuts, 30,253 minuts o 30,2527 minuts són possibles valors que pot prendre la variable. Altres exemples són el pes de les persones matriculades als gimnasos, l'edat exacta dels assistents habituals als congressos de matemàtiques o les tones que carreguen els camions d'una companyia de transport.

Nota

El conjunt de valors que pot prendre una variable aleatòria X s'anomena *rang de X*.

7.3.1. Variables aleatòries discretes

Com s'ha definit amb anterioritat, una variable aleatòria és discreta si únicament pot prendre un nombre finit de valors reals o, com a màxim, un conjunt numerable. Per tant, en aquest tipus de distribucions té sentit calcular la probabilitat que la variable aleatòria prenga un valor concret. Intuïtivament pareix clar, perquè en l'assignació de probabilitats s'ha de repartir 1 entre un conjunt finit o numerable de valors.

Funció de probabilitat

En les variables aleatòries discretes es defineix una funció, denominada *funció de probabilitat*, que associa a cada valor de la variable aleatòria la seua probabilitat. Per altra part, com que se sol representar per X (en majúscules) la variable aleatòria i per x (en minúscules) el valor que pren, es pot definir la funció de probabilitat en aquesta notació com: $f(x) = P(X = x)$.

Exemple 7

En l'exemple 6 es té que: X és una variable aleatòria discreta perquè només pot prendre un conjunt finit de valors distints: {2,3,4,5,6,7,8}.

A més, la funció de probabilitat seria la següent:

$$f(2) = P(X = 2) = \frac{1}{16} \qquad f(3) = P(X = 3) = \frac{1}{8} \qquad f(4) = P(X = 4) = \frac{3}{16}$$

$$f(5) = P(X = 5) = \frac{1}{4} \qquad f(6) = P(X = 6) = \frac{3}{16} \qquad f(7) = P(X = 7) = \frac{1}{8}$$

$$f(8) = P(X = 8) = \frac{1}{16}.$$

Gràficament les probabilitats es poden representar mitjançant un diagrama de barres (figura 1).

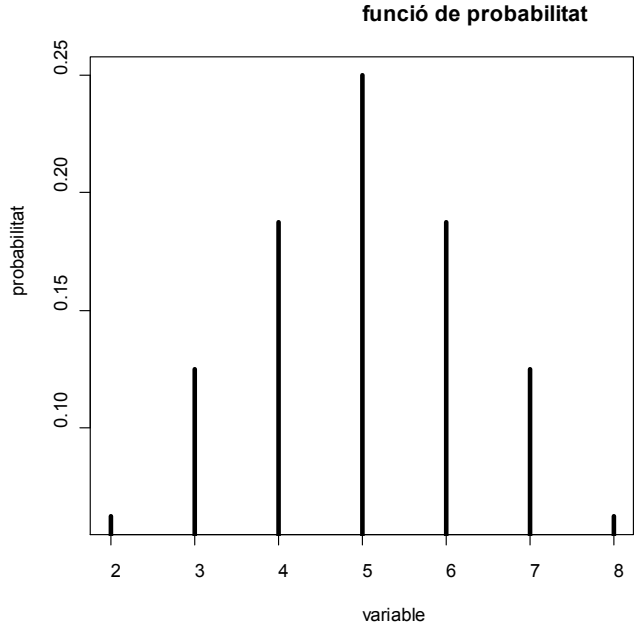


Figura 1

D'altra banda, $f(x)$ determina la distribució de probabilitats d'un experiment aleatori. És per això que ha de complir els axiomes de probabilitat, els quals es tradueixen en la nova notació de la manera següent:

Es consideren $\{x_1, .., x_n\}$ els diferents valors que pot prendre la variable aleatòria X. Llavors la funció de probabilitat ha de complir:

- $0 \leq f(x) \leq 1$ per a qualsevol x del conjunt $\{x_1 ... x_n\}$.
- $f(x_1) + f(x_2) + ... + f(x_n) = 1$

A més a més, si $A = \{x_2, x_3, x_k\}$; llavors $P(X \in A) = f(x_1) + f(x_2) + f(x_k)$.

La representació gràfica (figura 2) de la funció de probabilitat d'una distribució discreta és del tipus:

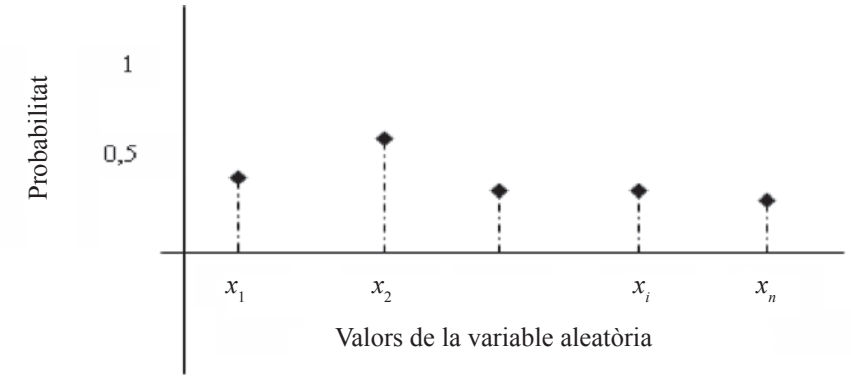


Figura 2

Nota

Cal assenyalar que les distribucions de probabilitat discretes es representen de la mateixa manera que les variables estadístiques en què les dades no s'agrupen. En tots dos casos s'empra un diagrama de barres. En les variables estadístiques es representa la freqüència relativa (o l'absoluta) i en les variables aleatòries discretes, la probabilitat.

Exemple 8

En l'exemple 6 es pot comprovar que sí que es compleixen els axiomes:

- $0 \leq f(x) \leq 1$ per a qualsevol valor que pren x entre els valors del rang de X: {2, 3, 4, 5, 6, 7, 8}.
- $f(2) + f(3) + f(4) + f(5) + f(6) + f(7) + f(8) =$

  $$= \frac{1}{16} + \frac{1}{8} + \frac{3}{16} + \frac{1}{4} + \frac{3}{16} + \frac{1}{8} + \frac{1}{16} = 1.$$

A més a més, la probabilitat de l'esdeveniment «sumar més de 5», és la probabilitat que la variable aleatòria X prenga els valors 6, 7 o 8.

Llavors $P(\text{sumar més de 5}) = P(X \in \{6, 7, 8\}) = f(6) + f(7) + f(8) =$

$$= \frac{3}{16} + \frac{1}{8} + \frac{1}{16} = \frac{3}{8}.$$

Funció de distribució

La funció de distribució d'una variable aleatòria té el mateix significat que la freqüència relativa acumulada de les distribucions estadístiques descriptives. Així, donada una distribució estadística discreta X en què a és un valor numèric, es defineix la funció de distribució $F(a)$ com la probabilitat que la variable aleatòria X no prenga un valor superior a a. És a dir:

$$F(a) = P(X \leq a) = \sum_{x \leq a} P(x) = \sum_{x \leq a} f(x).$$

Nota

Per exemple, si una variable aleatòria pren els valors {1,2,3,5}, llavors $F(4) = P(X = 3) + P(X = 2) + P(X = 1)$.

Propietats

- La funció de distribució es caracteritza per ser creixent, és a dir, si $x_1 \geq x_2 \rightarrow F(x_1) \geq F(x_2)$.

- $F(-\infty) = P(X \leq -\infty) = 0$, ja que cap valor de la variable aleatòria pot ser inferior a menys infinit.

- $F(+\infty) = P(X \leq +\infty) = 1$, ja que tots els valors de la variable aleatòria són inferiors a més infinit.

- $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) = \sum_{x>a}^{b} f(x)$. Per exemple, si una variable aleatòria pren els valors {1, 2, 3, 5}, llavors $P(2 < X \leq 5) = F(5) - F(2) = P(X = 5) + P(X = 4) + P(X = 3)$.

Exemple 9

Calcula la funció de distribució de l'exemple 6 i representa-la gràficament.

La que cal construir és una funció definida a trossos, ja que en cada interval comprès entre dos valors de la variable estadística la funció és constant. En aquest cas, l'expressió de la funció de distribució és:

$$F(x) = \begin{cases} 0 & si \quad x < 2 \\ 0{,}0625 & si \quad 2 \leq x < 3 \\ 0{,}1875 & si \quad 3 \leq x < 4 \\ 0{,}3750 & si \quad 4 \leq x < 5 \\ 0{,}6250 & si \quad 5 \leq x < 6 \\ 0{,}8125 & si \quad 6 \leq x < 7 \\ 0{,}9375 & si \quad 7 \leq x < 8 \\ 1 & si \quad 8 \leq x \end{cases}$$

S'han expressat les fraccions amb nombres decimals per comoditat. La representació gràfica d'aquesta funció (figura 3) és la següent:

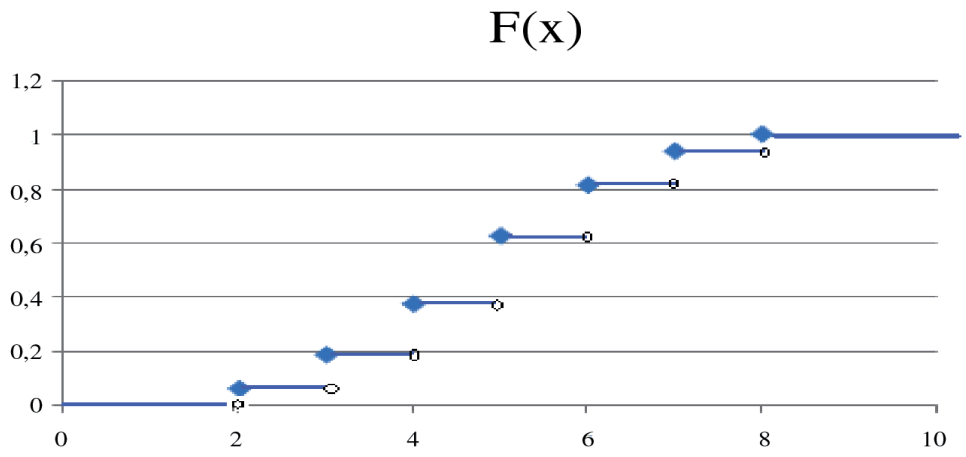


Figura 3

D'aquesta manera, és possible calcular probabilitats acumulades únicament a partir del gràfic o de la funció. Així, per exemple, la probabilitat que la variable X prenga un valor inferior o igual a 5 és $F(5) = 0{,}6250$, ja que 5 està en l' interval $5 \leq x < 6$. De la mateixa manera, la probabilitat que la funció prenga un valor inferior o igual a 4,5 és $F(4{,}5) = 0{,}3750$.

Esperança i variància d'una variable aleatòria discreta

De la mateixa manera que ocorria amb les variables estadístiques, en què es definien estadístics per a caracteritzar i, en certa manera, resumir la informació de les dades, en les variables aleatòries també és possible definir estadístics amb el mateix propòsit. En aquest cas sintetitzen la informació de la distribució de probabilitat de les variables aleatòries que representen.

Valor esperat o esperança matemàtica

Donada X una variable aleatòria, l'esperança matemàtica pot considerar-se com el valor esperat en realitzar l'experiment, el valor central de la distribució de probabilitats. Es calcula multiplicant cada valor de la variable aleatòria per la seua probabilitat.

Així, si X és una variable aleatòria discreta que pren els valors $\{x_1, x_2, ..., x_n\}$ i $f(x)$ és la funció de probabilitat associada, l'esperança de X:

$$\mu = E(X) = x_1 \cdot f(x_1) + x_2 \cdot f(x_2) + x_3 \cdot f(x_3) + ... + x_n f(x_n) = \sum_{i=1}^{n} x_i \cdot f(x_i).$$

De la definició de *valor esperat*, també es pot deduir una altra definició equivalent. Així, l'esperança d'una variable aleatòria discreta és la mitjana ponderada de tots els possibles resultats on els pesos són les probabilitats respectives dels resultats.

Exemple 10

En l'exemple 6 que s'està tractant al llarg d'aquest epígraf:

$$E(X) = 2 \cdot f(2) + 3 \cdot f(3) + 4 \cdot f(4) + 5 \cdot f(5) + 6 \cdot f(6) + 7 \cdot f(7) + 8 \cdot f(8) =$$

$$= 2 \cdot \frac{1}{16} + 3 \cdot \frac{1}{8} + 4 \cdot \frac{3}{16} + 5 \cdot \frac{1}{4} + 6 \cdot \frac{3}{16} + 7 \cdot \frac{1}{8} + 8 \cdot \frac{1}{16} = 5.$$

Per tant, el valor esperat quan es llancen dos daus i se sumen els resultats és 5.

Exemple 11

Un joc d'atzar molt senzill consiteix en el següent: es llança un dau equilibrat, si el resultat és {1, 2}, es guanyen un terç dels diners que s'han apostat en el joc. Si ix {3, 4} es guanyen la meitat dels diners que s'han apostat i si ix {5, 6} es perden tots els diners. Si una persona aposta 60 €, quin és el nombre d'euros que s'espera que guanye?

En aquest exemple, cal definir la variable aleatòria X = diners guanyats en el joc», a partir de l'experiment «llançar un dau. Així, els diferents valors que pot prendre X són {20, 30, –60} i la funció de probabilitat associada a la variable aleatòria X és:

$$f(20) = P(\textit{que el resultat del dau siga } 1 \textit{ o } 2) = \frac{1}{3}$$

$$f(30) = P(\textit{que el resultat del dau siga } 3 \textit{ o } 4) = \frac{1}{3}$$

$$f(-60) = P(\textit{que el resultat del dau siga } 5 \textit{ o } 6) = \frac{1}{3}$$

I, per tant, el valor esperat del guany és:

$$E(X) = 20 \cdot f(20) + 30 \cdot f(30) + (-60) \cdot f(60) =$$

$$= 20 \cdot \frac{1}{3} + 30 \cdot \frac{1}{3} - 60 \cdot \frac{1}{3} = \frac{-10}{3} = -3{,}33 \text{ €}.$$

És a dir, en aquest joc, el valor esperat és perdre 3,33 € jugant-ne 60.

Nota

Un joc d'atzar es diu que és just quan $E(X) = 0$.

Nota

Cal assenyalar que, d'una manera natural, es pot obtenir el model de probabilitat discret per a aquest joc.

Així, fixada la quantitat que s'hi vol jugar, anomenada C, els diferents valors que pot prendre la variable aleatòria són: $\left\{\frac{C}{3};\frac{C}{2};-C\right\}$, i la funció de probabilitat assignada és: $f\left(\frac{C}{3}\right)=\frac{1}{3};\ f\left(\frac{C}{2}\right)=\frac{1}{3};\ f(-C)=\frac{1}{3}$. I a partir de la funció de probabilitat es poden calcular els estadístics: $E(X)=\frac{C}{3}\cdot\frac{1}{3}+\frac{C}{2}\cdot\frac{1}{3}-C\cdot\frac{1}{3}=-\frac{C}{18}$.

Variància

Donada X una variable aleatòria, la variància pot considerar-se com una mesura de la dispersió dels valors respecte del valor esperat, en termes de probabilitat. És a dir, conceptualment és el mateix que el que es calcula en l'estadística descriptiva. S'obté cercant la mitjana aritmètica del quadrat de les desviacions respecte de l'esperança.

Així, si X és una variable aleatòria discreta que pren els valors $\{x_1, x_2, ..., x_n\}$ i $f(x)$ és la funció de probabilitat, llavors:

$$\text{Var}(X)=\sigma^2=(x_1-\mu)^2 f(x_1)+(x_2-\mu)^2 f(x_2)+...+(x_n-\mu)^2 f(x_n)=\sum_{i=1}^{n}(x_i-\mu)^2 f(x_i)\cdot$$

Nota

Es pot demostrar matemàticament que $\text{Var}(X) = E(X^2) - E(X)^2$. A més a més, de la mateixa manera que en les variables estadístiques, es calcula la desviació típica de la distribució de probabilitat com $\sigma=\sqrt{\sigma^2}$.

Exemple 12

Es calcularà la variància de la distribució de probabilitats del joc que apareix en l'exemple 11.

$$\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 f(x_i) = (x_1 - \mu)^2 f(x_1) + (x_2 - \mu)^2 f(x_2) + (x_3 - \mu)^2 f(x_3) =$$
$$= (20 - (-3{,}33))^2 \cdot \frac{1}{3} + (30 - (-3{,}33))^2 \cdot \frac{1}{3} + (-60 - (-3{,}33))^2 \cdot \frac{1}{3} = 1622{,}22.$$

Per a calcular la variància emprant l'altra equació, cal calcular primerament $E(X^2)$.

$$E(X^2) = \sum_{i=1}^{n} x_i^2 \cdot f(x_i) = 20^2 \cdot f(20) + 30^2 \cdot f(30) + (-60)^2 \cdot f(-60) =$$
$$= 400 \cdot \frac{1}{3} + 900 \cdot \frac{1}{3} + 3600 \cdot \frac{1}{3} = \frac{4900}{3} = 1633{,}3 \ €^2.$$

Per tant,

$$\sigma^2 = E(X^2) - E(X)^2 = 16333{,}33 - (-3{,}33)^2 = 1622{,}22.$$

I així, la desviació típica és 40,27 €, que és un valor molt alt per a la distribució de probabilitat.

7.3.2. Variables aleatòries contínues

Tal com s'ha comentat en el primer epígraf, una variable aleatòria és contínua si pot prendre qualsevol dels infinits valors compresos en un interval de la recta real. Per exemple, si es considera la variable X = altura d'una persona, X pot prendre qualsevol valor de l'interval (1,85, 1,86) –els valors 1,8501; 1,80502; 1,8053...; 1,85001; 1,85002...; 1,85011; 1,85012... en són uns quants exemples. Com pot intuir-se amb facilitat, pareix que existeixen més que infinits valors entre 1,85 i 1,86.

Consegüentment, no té sentit calcular la probabilitat que X prenga un valor concret, perquè n'hi ha infinits i aleshores la probabilitat és 0. Tanmateix, sí que té sentit el càlcul de la probabilitat que X prenga valors dins d'un interval $[a, b]$, és a dir, $P(X \in [a, b]) = P(a \leq X \leq b)$.

Ara bé, si en les distribucions discretes es definia la funció de probabilitat que permetia el càlcul de la probabilitat que X siga igual a un valor concret $f(x) = P(X = x)$, en les variables aleatòries contínues no es pot definir una funció d'aquest tipus. Es defineix, en canvi, per a cada model, una funció denominada *funció de densitat de probabilitat*, $f(t)$, la qual permet calcular la probabilitat que la variable aleatòria prenga valors dins d'un interval.

El concepte utilitzat per al càlcul de la probabilitat és el de *àrea*. Concretament, es defineix la probabilitat que X prenga valors dins de l'interval $[a, b]$ com l'àrea compresa entre la funció de densitat, l'eix horitzontal de les abscisses i els punts a i b. En el gràfic (figura 4), $P(X \in [0{,}5,1]) =$ àrea ombrejada $= \int_{0,5}^{1} f(t)dt$ (ja que en terminologia matemàtica l'àrea es representa mitjançant la integral).

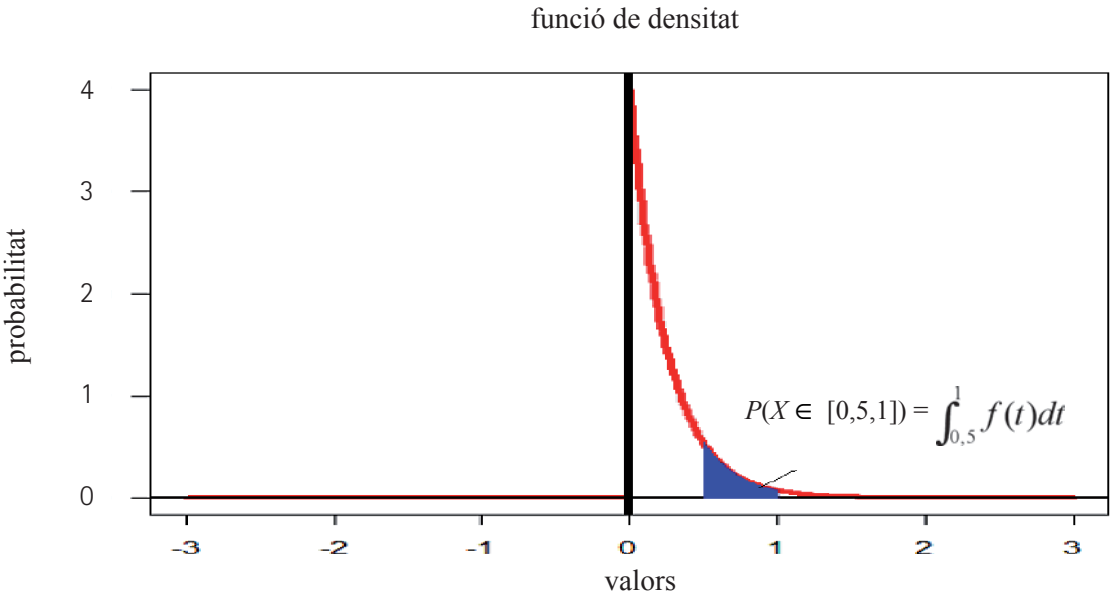


Figura 4

Nota

Cal recordar que en les variables estadístiques en què s'agrupaven les dades en intervals la representació gràfica de les dades es realitzava mitjançant els histogrames, les àrees dels quals representaven la freqüència de l'interval.

En les variables aleatòries contínues es representa la probabilitat emprant el mateix concepte de *àrea*. En aquest cas, però, cal conèixer la funció de densitat de probabilitat.

Com s'ha comentat, amb la funció de densitat no és possible calcular directament la probabilitat, tanmateix sí que permet calcular-la emprant el concepte *integral*.

D'altra banda, com que la funció de densitat permet determinar la distribució de probabilitats d'un experiment aleatori, és necessari que complisca els axiomes de probabilitat, els quals es tradueixen en:

- $f(x) \geq 0$ (la funció és positiva perquè l'àrea és positiva) (figura 5, esquerra).
- $\int_{-\infty}^{+\infty} f(t)dt = 1$ (l'àrea total ha de ser 1) (figura 5, dreta).

A més a més, si $A = [a, b]$, $P(X \in A) = P(X \in [a, b]) = P(a \leq X \leq b) = \int_a^b f(t)dt$

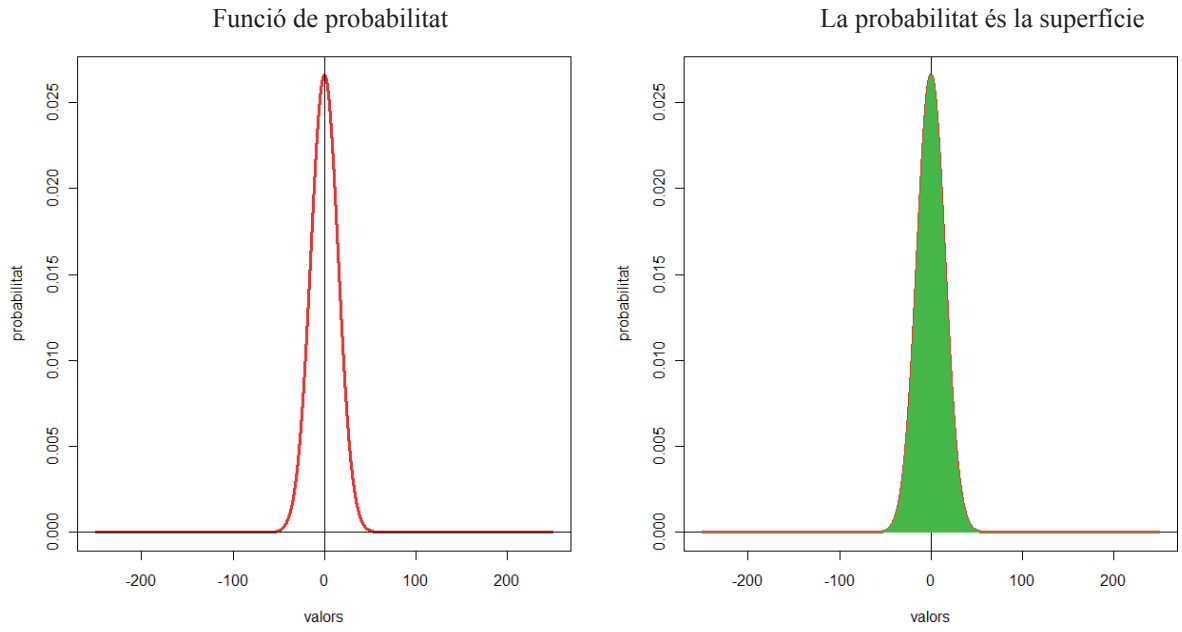


Figura 5

Exemple 13

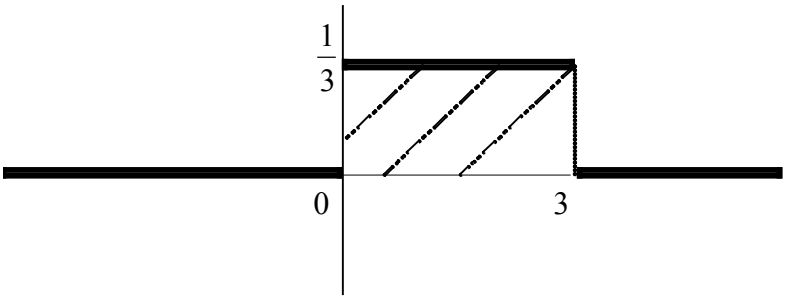
Es considera X = instant de temps en què s'espera una trucada de telèfon entre les 0 h i les 3 h. Se suposa que la trucada es produeix aleatòriament dins de l'interval de temps.

En aquest cas, com que l'instant de temps és aleatori, la probabilitat que la trucada es produïsca entre les 0.30 h i la 1.30 h o entre les 0.55 h i la 1.55 h o entre les 0.32 h i la 1.32 h ..., és la mateixa. És a dir, la probabilitat únicament depèn de l'amplària de l'interval, i intervals de temps de la mateixa amplària tenen la mateixa probabilitat de rebre la trucada. És per això que la funció de quantia ha de ser constant al llarg de tot el domini de definició de la variable aleatòria ($f(t) = k$ per a qualsevol t de l'interval [0,3]). A més a més, l'àrea compresa entre la funció de densitat, l'eix d'abscisses, els valors 0:00 i 3:00, i la funció de quantia, és 1. D'aquesta darrera afirmació es pot obtenir la constant k:

$$1 = \text{Àrea total} = \int_{-\infty}^{+\infty} f(t)dt = \int_0^3 f(t)dt = \int_0^3 kdt = (3-0)\cdot k \text{ i resolent l'equació, } k = \frac{1}{3}.$$

D'aquesta manera s'obté la funció de densitat de l'experiment:

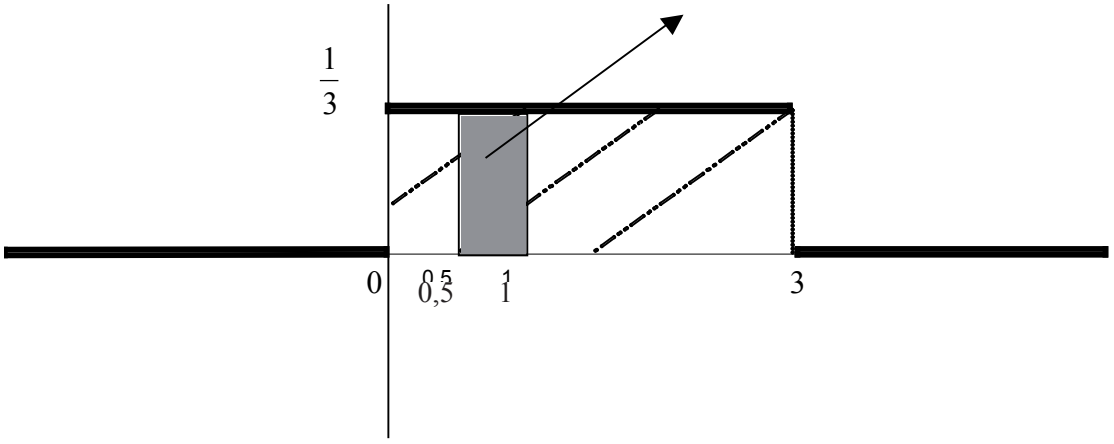
$$f(t) = \begin{cases} \frac{1}{3} & \text{si} \quad 0 \le t \le 3 \\ 0 & \text{si} \quad t < 0 \ \ o \ t > 3 \end{cases}$$



A partir d'aquesta funció, és possible calcular, per exemple, la probabilitat que la trucada es produïsca entre les 0:50 h i la 1:00 h.

Àrea compresa entre $f(t)$, eix d'abscisses, $x = 0{,}5$ i $x = 1$.

És a dir, $0{,}5 \cdot \frac{1}{3} = \frac{1}{6}$ (emprant la fórmula de l'àrea d'un rectangle)



Així doncs, $P(0{,}5 \le X \le 1) = \int_{0,5}^{1} \frac{1}{3} dt = 0{,}5 \cdot \frac{1}{3} = \frac{1}{6}$.

Anàlogament, la $P(0{,}75 \le X \le 1{,}25) = \int_{0,75}^{1,25} \frac{1}{3} dt = 0{,}5 \cdot \frac{1}{3} = \frac{1}{6}$.

Nota

- Com s'ha pogut comprovar, si la funció de densitat és constant, la probabilitat que la variable aleatòria prenga un valor dins de l'interval $[c, d]$, a més a més de calcular-se resolent la integral, també es pot calcular aplicant la fórmula de l'àrea d'un rectangle, de base l'amplària de l'interval $(c - d)$ i d'altura, el valor constant de la funció. Així, en l'exemple, $P(0{,}35 \leq X \leq 1{,}25) = 0{,}9 \cdot \frac{1}{3} = 0{,}3$.
- També es pot comprovar que la probabilitat únicament depèn de l'amplària de l'interval.

Funció de distribució

La funció de distribució d'una variable aleatòria contínua té el mateix significat que la de la funció de la variable aleatòria discreta. El que varia és la manera de calcular la probabilitat demanada. Així, donada una distribució estadística contínua discreta X en què a és un valor numèric, es defineix la funció de distribució $F(a)$ com la probabilitat que la variable aleatòria X no prenga un valor superior a a. És a dir:

$$F(a) = P(X \leq a) = \int_{-\infty}^{a} f(t)dt.$$

El gràfic següent mostra la funció de densitat i la funció de distribució per a una distribució de probabilitat contínua. Com s'observa, la superfície marcada en el gràfic de la funció de densitat $P(X \leq 100)$ és el valor que correspon a 100 en el gràfic de la funció de distribució.

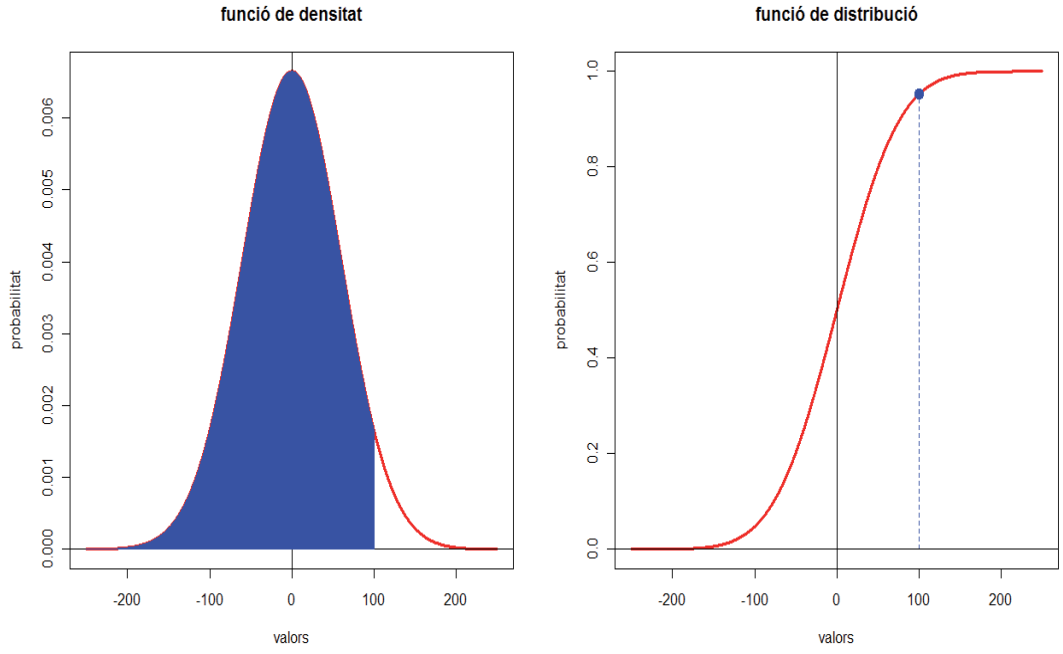


Figura 6

Propietats

- La funció de distribució es caracteritza per ser creixent, és a dir, si $x_1 \geq x_2 \rightarrow F(x_1) \geq F(x_2)$.
- $F(-\infty) = P(X \leq -\infty) = 0$, ja que cap valor de la variable aleatòria pot ser inferiors a menys infinit.
- $F(+\infty) = P(X \leq +\infty) = 1$, ja que tots els valors de la variable aleatòria són inferiors a més infinit.

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = \int_{a}^{b} f(x)dx$$

Les dues primeres propietats es poden comprovar intuïtivament en els gràfics anteriors. Aquesta darrera propietat és molt important i de molta utilitat a l'hora de fer càlculs de probabilitats. Els gràfics següents en faciliten la comprensió (figura 7).

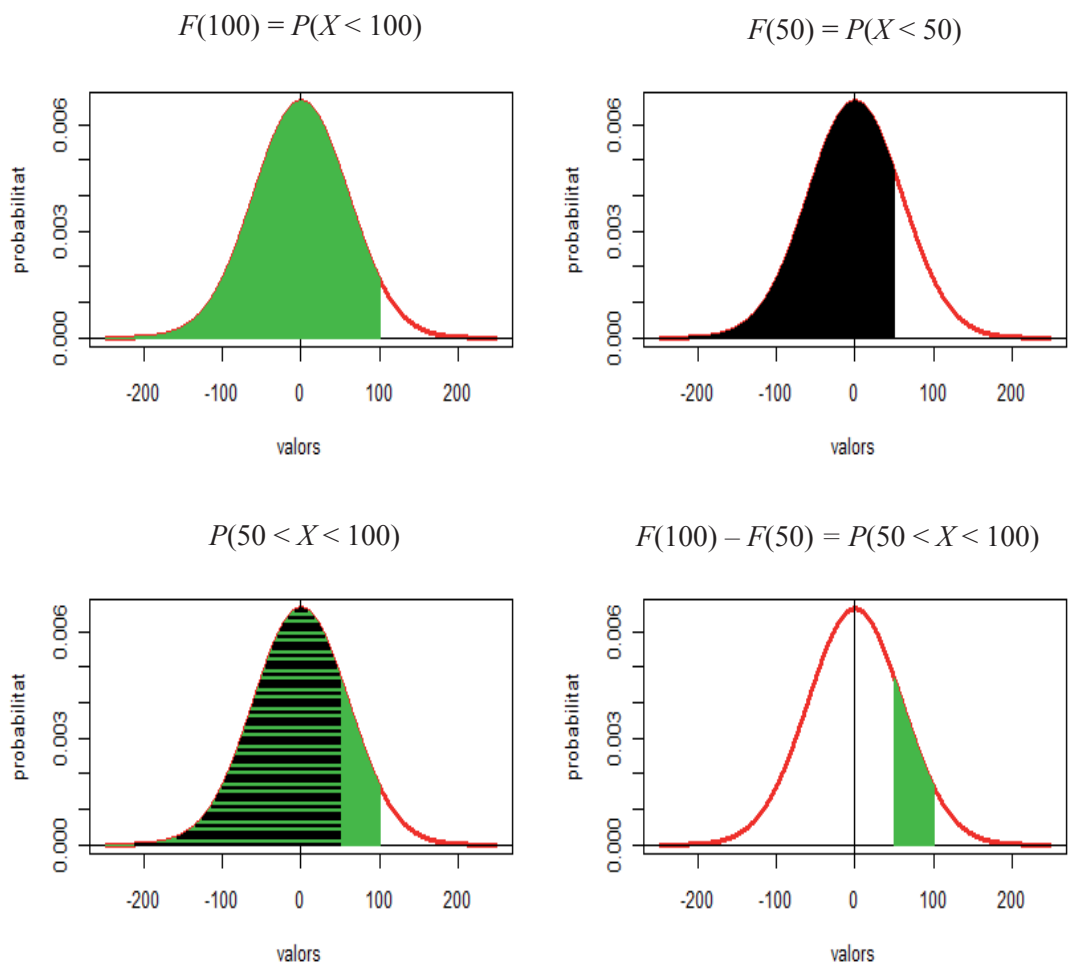


Figura 7

Exemple 14

Calcula la funció de distribució de l'exemple 18 i representa-la gràficament. Calcula també la probabilitat $P(0{,}75 \leq X \leq 1{,}5)$.

En les variables aleatòries contínues trobar la funció de distribució és molt més complex que en les discretes, ja que cal calcular la integral de la funció de densitat. Com que la funció de densitat de l'exemple 14 és:

$$f(t) = \begin{cases} \frac{1}{3} & \text{si} \quad 0 \leq t \leq 3 \\ 0 & \text{si} \quad t < 0 \ \ o \ \ t > 3 \end{cases}$$

llavors, fent la integral d'aquesta funció, s'obté la funció de distribució:

$$F(t) = \begin{cases} \frac{1}{3}t & \text{si} \quad 0 \leq t \leq 3 \\ 0 & \text{si} \quad t < 0 \ \ o \ \ t > 3 \end{cases}$$

I la seua representació gràfica és una recta (figura 8).

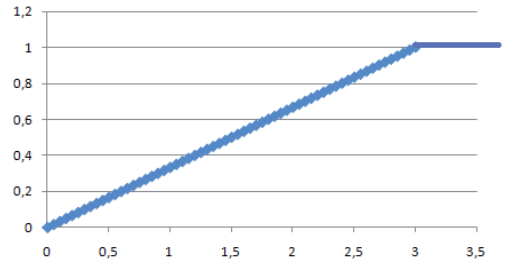


Figura 8

Amb la funció de distribució és molt fàcil calcular probabilitats acumulades. Així, la probabilitat que X siga inferior o igual a 1,5 és $\frac{1}{3} \cdot 1{,}5 = 0{,}5$, que gràficament equivaldria al valor de la funció en 1,5 (figura 9).

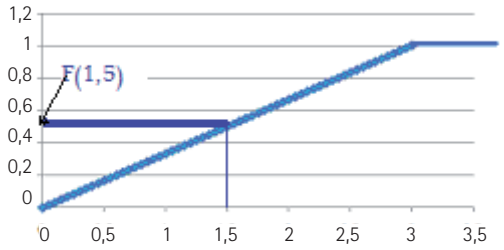


Figura 9

A més a més, la funció de distribució també permet calcular la probabilitat d'intervals. Així:

$$P(0{,}75 \leq X \leq 1{,}5) = F(1{,}5) - F(0{,}75) = \frac{1}{3}\cdot 1{,}5 - \frac{1}{3}\cdot 0{,}75 = 0{,}25.$$

Esperança i variància d'una variable aleatòria contínua

De la mateixa manera que ocorria amb les variables aleatòries discretes, es poden definir els estadístics «esperança matemàtica» i «variància» per a les variables contínues. Cal dir que, malgrat tenir significats equivalents en tots dos tipus de variables aleatòries, en les variables contínues són en molts casos aquests paràmetres els que caracteritzen completament la distribució. Fins al punt de ser necessaris per a poder definir la funció de densitat de probabilitat.

Valor esperat o esperança matemàtica

Si X és una variable aleatòria contínua i $f(t)$ és la funció de quantia associada, l'esperança de X es calcula així:

$$\mu = E(X) = \int_{-\infty}^{\infty} t \cdot f(t)\,dt.$$

Variància

Siga X una variable aleatòria contínua i $f(t)$ la funció de quantia que té associada, la variància de X es calcula de la manera següent:

$\mathrm{Var}(X) = \sigma^2 = \int_{-\infty}^{\infty} (t-\mu)^2 f(t)\,dt$ o l'expressió equivalent, $\mathrm{Var}(X) = E(X^2) - E(X)^2$.

Exemple 15

Calculem l'esperança matemàtica i la variància de l'exemple 14. Si X és la variable aleatòria i $f(t)$ és la seua funció de densitat de probabilitat:

$$f(t) = \begin{cases} \frac{1}{3} & \text{si} \quad 0 \leq t \leq 3 \\ 0 & \text{si} \quad t < 0 \ \ o \ \ t > 3 \end{cases}$$

Llavors $E(X)=\int_{-\infty}^{\infty} t\cdot f(t)dt=\int_{0}^{3} t\cdot\frac{1}{3}dt=\left[\frac{t^2}{6}\right]_0^3=\frac{9}{6}-0=\frac{3}{2}=1{,}5$. Aleshores, l'instant de temps en què s'espera que es produïsca la trucada és a les 1,5 hores.

I la $\text{Var}(X)=E(X^2)-E(X)^2=\int_0^3 t^2\,\frac{1}{3}dt-\left(\frac{3}{2}\right)^2=\left[\frac{t^3}{9}\right]_0^3-\left(\frac{3}{2}\right)^2=\frac{27}{9}-\left(\frac{9}{4}\right)=\frac{3}{4}$.

7.4. Distribució conjunta de dues variables aleatòries

En el capítol cinquè es van introduir les variables estadístiques bidimensionals que sorgien en considerar dues característiques numèriques conjuntament. De la mateixa manera és possible que siga d'interès l'estudi conjunt de dues variables aleatòries. Per exemple, calcular la probabilitat que es venguen 4.000 productes que fabrica una determinada empresa a un preu de 25€ la unitat o conèixer la probabilitat que un treballador d'un país té d'arribar a guanyar un sou determinat si domina quatre idiomes, poden ser de molta utilitat.

Així, si es considera un experiment aleatori i a cada esdeveniment elemental s'associen dos valors numèrics (x, y) el que s'obté és una variable aleatòria bidimensional (X,Y). És evident que aquesta associació ha de tenir sentit, és a dir, no és coherent associar la mateixa característica per al primer valor (x) que per al segon (y), ja que en aquest cas únicament es necessitaria una variable aleatòria. En conseqüència, els valors que prenen les x i les y representen aspectes diferents i llavors es pot parlar de dues variables aleatòries, X i Y. La figura 10 reflecteix gràficament el concepte de *variable aleatòria bidimensional*. Així, el gràfic mostra la distribució espacial en una determinada regió de les persones que fumen (i les que no), així com de les persones que han desenvolupat la malaltia del càncer (i les que no). S'hi empren: el color per a determinar les persones que fumen (negre, fumadors; blau, no-fumadors) i la grandària dels punts per a mostrar les persones malaltes (punts grossos, malalts; resta de punts, no-malalts).

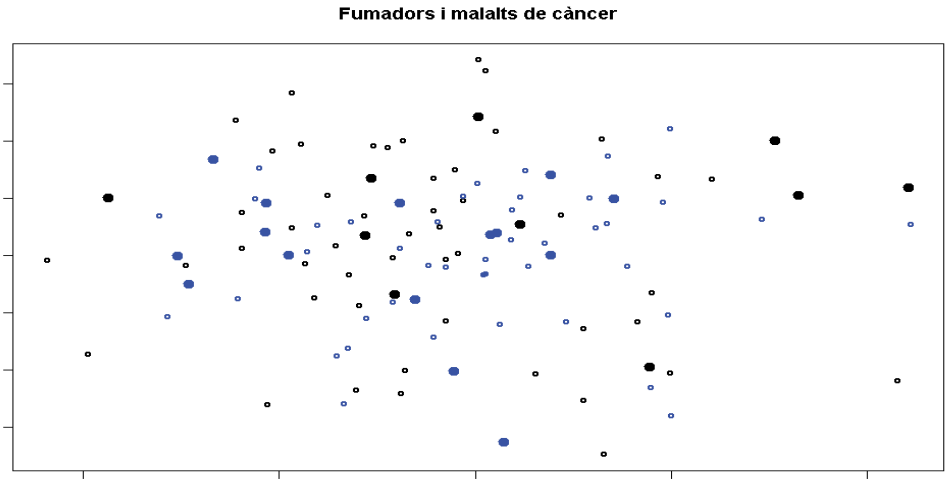


Figura 10

D'aquesta manera l'experiment aleatori que consisteix a extraure una persona del grup i saber si fuma i si ha desenvolupat càncer té per espai mostral E = {fuma i té càncer, fuma i no té càncer, no fuma i té càncer, no fuma i no té càncer}. Per a aconseguir la variable aleatòria, podem definir dues variables aleatòries X i Y de manera que X valga 0 si la persona no fuma i 1 en l'altre cas, i Y valga 0 si la persona no està malalta i 1 si ho està.

Amb ambdues variables obtenim la variable bidimensional (X, Y), que pot prendre els valors que apareixen en la taula de la (figura 11) segons els resultats que s'han obtingut en l'experiment. És a dir, a l'esdeveniment elemental «fuma i té càncer», s'hi associa (1, 1); a «fuma i no té càncer», (1, 0)... En conseqüència, els valors que pot prendre la variable aleatòria (X, Y) són {(0, 0), (1, 0), (0, 1), (1, 1)}.

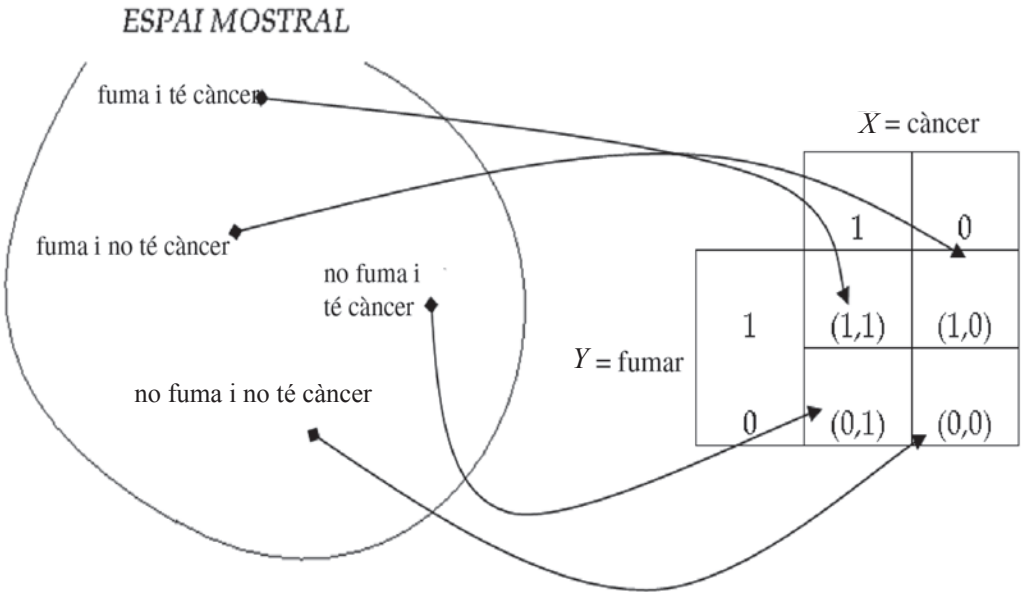


Figura 11

De la mateixa manera que per a una variable aleatòria, la probabilitat de cada parell (x, y) està determinada per la probabilitat de l'esdeveniment que representa. Així, en l'exemple de la figura 10 que estem analitzant hi ha:

- 45 observacions de persones que fumen i no tenen càncer.
- 15 observacions de persones que fumen i tenen càncer.
- 50 observacions de persones que no fumen i no tenen càncer.
- 10 observacions de persones que no fumen i tenen càncer.

I, per tant, les probabilitats són:

- $P((X, Y) = (0, 0)) = P(X = 0, Y = 0) = P(0, 0) = \frac{50}{120}$;
- $P((X, Y) = (1, 0)) = P(X = 1, \mathrm{Y} = 0) = P(1, 0) = \frac{45}{120}$;
- $P((X, Y) = (0, 1)) = P(X = 0, Y = 1) = P(0, 1) = \frac{10}{120}$;
- $P((X, Y) = (1, 1)) = P(X = 1, Y = 1) = P(1, 1) = \frac{15}{120}$.

És a dir, $P(x,y) = P(X = x, \mathrm{Y} = y)$.

Aquesta informació també sol estar representada en una taula com la següent:

		X	
	Prob.	0	1
Y	0	$\frac{50}{120}$	$\frac{45}{120}$
	1	$\frac{10}{120}$	$\frac{15}{120}$

Taula 7

En aquest exemple emprat per a introduir el concepte *variable aleatòria bidimensional* s'ha considerat que ambdues variables que la conformen són discretes. No obstant això, aquest fet no sempre és així.

D'altra banda, com ja s'ha comentat, cada una d'aquestes variables té entitat per si sola i, per tant, es pot fer l'estudi de la variable aleatòria de la mateixa manera que s'ha vist per a les variables aleatòries unidimensionals. Aquests tipus de variables s'anomenen *variables aleatòries marginals*, per analogia amb les variables estadístiques. En el cas que s'està considerant n'hi ha dues: X = fumador i Y = malalt de càncer.

7.4.1. Variables aleatòries bidimensionals discretes i contínues

En les variables aleatòries unidimensionals es pot distingir entre *variables aleatòries discretes* i *contínues*. En les variables aleatòries bidimensionals aquesta distinció també es realitza. Així, es diu que una variable (X, Y) és discreta si els valors possibles de (X, Y) són finits o infinits numerables. L'exemple que mostra la figura 10 és una variable aleatòria bidimensional, ja que únicament pren quatre possibles valors.

Tanmateix, (X, Y) és una variable aleatòria bidimensional contínua si pot prendre qualsevol valor comprès en un interval de dues dimensions (per exemple (1, 0) x (2,3) és un interval bidimensional i el formen tots els parells en què el primer terme està comprès entre 0 i 1, i el segon terme entre 2 i 3). Així, elements d'aquest interval són (0,75, 2,25), (0,7569, 2,0056), etc. La variable (X, Y) on la primera component és l'edat de les persones i la segona el temps treballat al llarg de la seua vida en un determinat país, és contínua.

Nota

En algunes ocasions és possible trobar variables aleatòries bidimensionals en què una component siga contínua i l'altra, discreta. Per exemple, si (X, Y) representa l'edat de les persones i el nombre d'empreses on han treballat. Cal dir, però, que en aquest text no es tractaran aquest tipus de variables.

Per altra part, tal com ocorre amb les variables aleatòries unidimensionals, si els resultats de l'experiment són numèrics, llavors les associacions entre nombres i esdeveniments elementals són directes.

Per a descriure la distribució de probabilitats de (X, Y) es procedeix d'una manera semblant a la del cas unidimensional.

Variables aleatòries bidimensionals discretes

Es considera (X, Y) una variable aleatòria bidimensional discreta. A cada possible resultat (x_i, y_j) s'associa un nombre $p(x_i, y_j)$ que representa $P(X = x_i, Y = y_j)$, de manera que satisfaça els axiomes de probabilitat (de la manera semblant a les variables unidimensionals):

a) $p(x_i, y_j) \geq 0$ per a tots els parells (x_i, y_j)

b) $\sum_{x_i} \sum_{y_j} p(x_i, y_j) = 1.$

La funció p definida per a tots els valors (x_i, y_j) del recorregut de (X, Y) s'anomena *funció de probabilitat de (X, Y)* i determina la distribució de probabilitat de la variable. A més a més, per a calcular la probabilitat d'un subconjunt A del recorregut cal sumar les probabilitats de cadascun dels valors que el formen:

$$P(A) = \sum_{(x_i, y_j) de\ A} p(x_i, y_j).$$

D'altra banda, és possible construir una gràfica de $p(x, y)$ com es mostra en la figura 12. Cada probabilitat pot representar-se mitjançant uns eixos tridimensionals, on la base, la constituixen els valors de les variables X i Y i l'altura, la funció de probabilitat p. Cal remarcar que aquest gràfic és molt semblant al que es va comentar en el capítol cinquè, referent a les variables estadístiques bidimensionals.

Variable aleatòria discreta

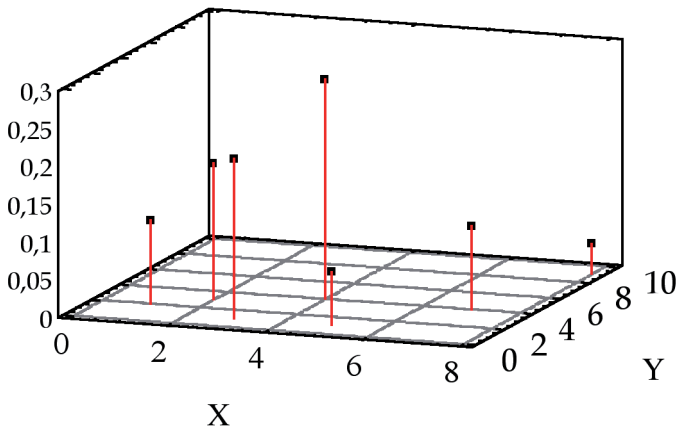


Figura 12

Nota

Aquesta definició es pot generalitzar per a un nombre indeterminat de variables aleatòries. Així, es pot parlar de *variables aleatòries n-dimensionals* $(X_1,...,X_n)$ i de la seua funció de probabilitat conjunta $p(x_1,...,x_n) = P(X_1 = x_1...,X_n = x_n)$, la qual també compleix els axiomes de probabilitat:

a) $p(x_1,...,x_n) \geq 0$ per a tots els parells $(x_1,...,x_n)$ de R de $(X_1,...,X_n)$

b) $\sum_{x_1}...\sum_{x_n} p(x_1,...,x_n) = 1.$

Exemple 16

Una empresa està intentant introduir un producte en Noia, una ciutat ubicada a la província de la Corunya. Per a fer-ho, ha encarregat a una altra empresa que estudie quina relació hi ha entre el nombre de fills menors de 10 anys que tenen les parelles i la predisposició a comprar un determinat producte. S'ha arribat als resultats següents:

- El 10% de les parelles tenen 0 fills i estan disposades a comprar el producte. El 25% de les parelles tenen 0 fills i no estan disposades a comprar el producte.
- El 20% de les parelles tenen 1 fill i estan disposades a comprar el producte. El 25% de les parelles tenen 1 fill i no estan disposades a comprar el producte.
- El 8% de les parelles tenen 2 fills i estan disposades a comprar el producte. El 9% de les parelles tenen 2 fills i no estan disposades a comprar el producte.
- El 2% de les parelles tenen 3 fills i estan disposades a comprar el producte. L'1% de les parelles tenen 3 fills i no estan disposades a comprar el producte.

Si seleccionem a l'atzar una parella de la ciutat, quina és la probabilitat que tinga dos fills i que estiga disposada a comprar el producte?

Per a contestar la pregunta, en primer lloc es definirà la variable aleatòria conjunta i la seua distribució de probabilitat.

Així, l'espai mostral d'aquest experiment està format per totes les parelles del tipus (*x*, sí) o (*x*, no) on *x* representa el nombre de fills que tenen les parelles (0, 1, 2 i 3). És clar que la primera component representa una variable aleatòria *X* = nombre de fills de la parella, i la segona component indica si la família està interessada en la compra del producte. Com que aquesta segona component no és una variable aleatòria, cal associar nombres a cada possible valor. Així, es pot associar el número 1 a estar interessat a comprar el producte i el 0 a no estar-hi interessat. Llavors s'obté la variable aleatòria bidimensional (*X*, *Y*), la qual té per rang {(0, 1); (1, 1); (2, 1); (3, 1); (0, 0); (1, 0); (2, 0); (3, 0)}.

La distribució de probabilitats, tenint en compte els percentatges que apareixien abans, és la que es mostra en la taula següent:

	X = nombre de fills per família			
Y = decisió de compra	0	1	2	3
0	0,25	0,25	0,09	0,01
1	0,1	0,2	0,08	0,02

Taula 8

Per altra banda, el que es demana és la probabilitat que tinga dos fills i estiga disposada a comprar el producte, la qual cosa és equivalent a:

$$P(2, 1) = P(X = 2, Y = 1) = 0{,}08.$$

Variables aleatòries bidimensionals contínues

Es considera (X, Y) una variable aleatòria bidimensional contínua que pren tots els valors possibles dins d'una regió R de dues dimensions (R és un subconjunt d'un interval de dues dimensions). El conjunt format per tots els possibles valors que pot prendre la variable aleatòria s'anomena *recorregut* o *rang*.

Pel mateix raonament que en les variables aleatòries unidimensionals, en les bidimensionals no té sentit parlar de la probabilitat que la variable aleatòria prenga un valor concret (un parell concret (x_i, y_j)), de fet aquesta probabilitat és 0. Per contra, és evident que hi haurà intervals bidimensionals més probables que d'altres i per això sorgeix la necessitat de definir una funció *f(x, y)* que ho reflectisca. Aquesta funció s'anomena *funció de densitat de probabilitat conjunta f*, i permet calcular la probabilitat d'un interval bidimensional mitjançant el càlcul integral que, si en considerar una dimensió representa l'àrea, en considerar-ne dues representa el volum.

D'altra banda, aquesta funció ha de complir les exigències definides en els axiomes de probabilitat:

a) $f(x, y) \geq 0$ per a tots els parells (x, y) de R de (X, Y),

b) $\int_R f(x, y)dxdy = 1.$

La segona condició significa que el volum total, el qual representa la probabilitat total, és 1.

Per altra part, és possible construir el gràfic que representa la funció de densitat de probabilitat conjunta. Així, a cada parell (*x*, *y*) del recorregut s'associa, a l'eix vertical, una altura igual al valor *f(x, y)*. La representació gràfica és òbviament una superfície, ja que tant *X* com *Y* són contínues. Cal fer notar que el volum que encercla la superfície i el pla *XY* es calcula segons la integral doble abans esmentada i és 1 (2a exigència dels axiomes de probabilitat.) La figura 13 mostra un exemple de la distribució de probabilitats d'una variable aleatòria bidimensional.

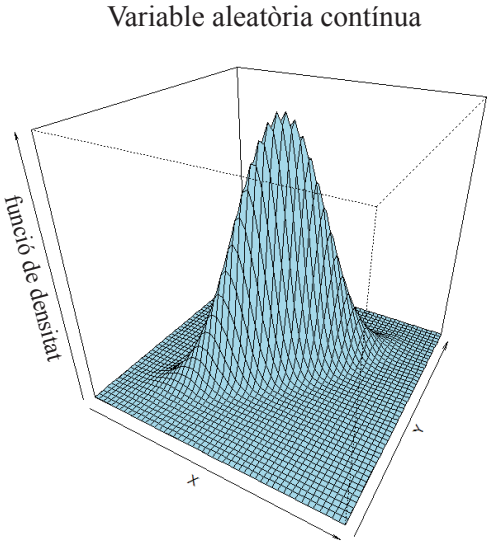


Figura 13

A més a més, per a calcular la probabilitat d'un subconjunt *A* del recorregut cal trobar el volum encerclat entre *A* i la funció de densitat (figura 14). Per a fer-ho s'ha de resoldre la integral doble sobre l'esmentat subconjunt:

$$P(A) = \int_A f(x,y)dxdy$$

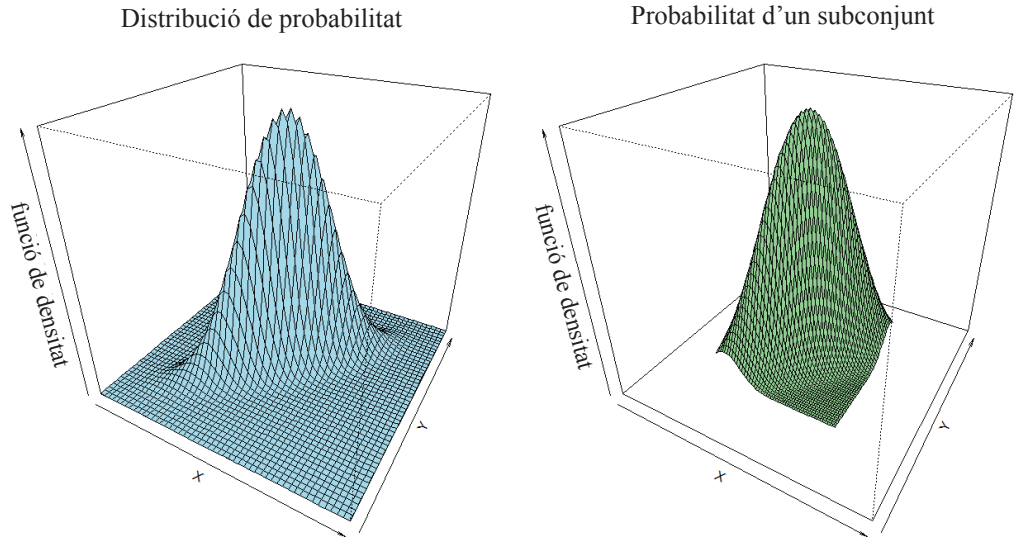


Figura 14

Nota

De la mateixa manera que en les variables discretes, la definició de *varible aleatòria bidimensional contínua* es pot generalitzar per a un nombre indeterminat de variables aleatòries contínues. Així, es pot parlar de *variables aleatòries n-dimensionals contínues* $(X_1,...,X_n)$ i de la seua funció de densitat de probabilitat conjunta $f(x_1,...,x_n)$, la qual pren valors més grans en aquelles regions més probables, i compleix els axiomes de probabilitat següents:

a) $f(x_1,...,x_n) \geq 0$ per a tots els parells $(x_1,...,x_n)$ de R de $(X_1,...,X_n)$,

b) $\int_R \cdots \int \cdots \int f(x_1,\ldots,x_n)dx_1 \ldots dx_n = 1$,

on R és el rang de la variable aleatòria multidimensional contínua.

Exemple 17

Un equip d'investigació intentà conèixer la població d'una determinada regió atès que havien descobert que una gran part d'aquesta tenia un ingressos econòmics alts. Després de molt de temps d'investigació van arribar a la conclusió que les variables que influïen en l'èxit econòmic eren el coeficient intel·lectual i la cultura dels individus. Així que decidiren estudiar aquestes dues variables en la població més exhaustivament. El primer que van fer fou *matematitzar* les dues característiques que estaven considerant. És a dir, van aconseguir assignar a cada persona un nombre de l'interval [0, 1] per a determinar el seu coeficient intel·lectual i un nombre de l'interval [0, 2] per a determinar-ne la cultura. Així mateix, designant X = coeficient intel·lectual i Y = grau de cultura, hipotetitzaren una funció de densitat de probabilitat conjunta per a la variable (X, Y). Aquesta funció és la següent:

$$f(x,y) = x^2 + \frac{x \cdot y}{3}.$$

La figura 15 mostra la representació gràfica de la funció de densitat, la qual permet deduir que en aquesta regió la probabilitat de trobar persones amb valors de la X i de la Y propers a l'1 i al 2, respectivament, és prou alta.

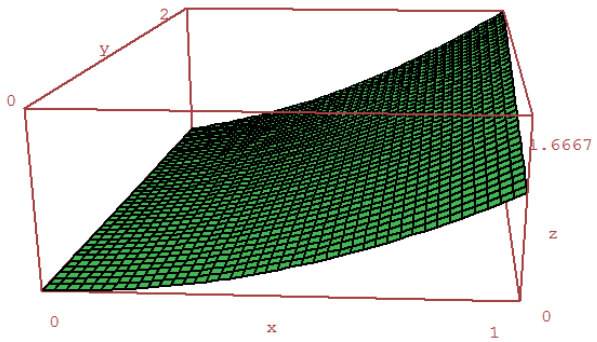


Figura 15

7.4.2. Distribucions marginals i condicionades

Donada una variable aleatòria bidimensional (X, Y), és possible que siga interessant l'estudi de la distribució de probabilitats d'una de les variables X o Y, o tal vegada l'estudi d'una de les variables aleatòries quan l'altra pren un (o més d'un) valor concret. Així, per exemple, si s'està estudiant la variable aleatòria bidimensional X = edat i Y = sou, pot ser que ens interesse estudiar la distribució de probabilitats de la variable «sou», o tal vegada necessitarem conèixer com es distribueix l'edat de les persones que tenen un sou superior a 1.600 €. Aquests casos responen a les distribucions marginal i condicionada, respectivament. Les distribucions marginals es representen per X i Y, i les condicionades per $X/(Y = y_i)$. Així mateix, es pot parlar de *funcions de probabilitat marginals de X i de Y*, en el cas de variables bidimensionals discretes, i de funcions de densitat de probabilitat marginals de X i de Y, si s'està considerant una variable bidimensional contínua.

Cal dir que aquest concepte de *distribució marginal* també es generalitza per a variables aleatòries n-dimensionals d'una manera natural.

Exemple 18

Prenent com a referència l'exemple 27, on X = nombre de fills, Y = decisió de compra, i la funció de probabilitat conjunta donada per la taula, calcula la funció de probabilitat condicionada de cadascuna de les variables. Calcula també la funció de probabilitat de X que condiciona que $Y = 1$.

	X = nombre de fills per família			
Y = decisió de compra	0	1	2	3
0	0,25	0,25	0,09	0,01
1	0,1	0,2	0,08	0,02

La funció de probabilitat marginal de X:

$$P(X = 0) = P(X = 0, Y = 0) + P(X = 0, Y = 1) = 0{,}25 + 0{,}10 = 0{,}35$$
$$P(X = 1) = P(X = 1, Y = 0) + P(X = 1, Y = 1) = 0{,}25 + 0{,}20 = 0{,}45$$
$$P(X = 2) = P(X = 2, Y = 0) + P(X = 2, Y = 1) = 0{,}09 + 0{,}08 = 0{,}17$$
$$P(X = 3) = P(X = 3, Y = 0) + P(X = 3, Y = 1) = 0{,}01 + 0{,}02 = 0{,}03.$$

La funció de probabilitat marginal de Y:

$$P(Y = 0) = P(X = 0, Y = 0) + P(X = 1, Y = 0) + P(X = 2, Y = 0) + P(X = 3, Y = 0)$$
$$= 0{,}25 + 0{,}25 + 0{,}09 + 0{,}01 = 0{,}6.$$

$$P(Y=1) = P(X=0, Y=1) + P(X=1, Y=1) + P(X=2, Y=1) + P(X=3, Y=1)$$
$$= 0{,}2 + 0{,}1 + 0{,}08 + 0{,}02 = 0{,}4.$$

La funció de probabilitat de X condiciona que $Y = 1$:

$$P(X=0/Y=1) = P(X=0, Y=1) = 0{,}10 = 0{,}10$$
$$P(X=1/Y=1) = P(X=1, Y=1) = 0{,}20 = 0{,}20$$
$$P(X=2/Y=1) = P(X=2, Y=1) = 0{,}08 = 0{,}08$$
$$P(X=3/Y=1) = P(X=3, Y=1) = 0{,}02 = 0{,}02.$$

Cal notar que:

$$P(X=0/Y=1) + P(X=1/Y=1) + P(X=2/Y=1) + P(X=3/Y=1) = P(Y=1).$$

7.4.3. Variables aleatòries bidimensionals independents

Al capítol sisè es van introduir els conceptes *esdeveniments independents* i *experiments independents*. Així, es deia que dos esdeveniments són independents quan l'ocurrència de l'un no influeix en l'ocurrència de l'altre. És a dir, A i B són dos esdeveniments independents si $P(A/B) = P(A)$, o l'expressió equivalent segons el teorema de caracterització: A i B són dos esdeveniments independents si $P(A \cap B) = P(A) \cdot P(B)$.

És evident que aquest concepte de *independència* també té la seua «versió» en les variables aleatòries, ja que aquestes no són res més que una altra manera d'expressar els esdeveniments. Així, si l'experiment consisteix a anotar, d'una banda el nombre de trucades de telèfon que rep una centraleta al llarg de les 5 hores de feina del matí d'una companyia a Madrid –nombre que anomenarem X–, i de l'altra, fer el mateix recompte però per a una centraleta a Barcelona –nombre que anomenarem Y–, les variables X i Y són intuïtivament independents. El nombre de telefonades que es reben en un lloc és independent del que es reben en l'altre. En conseqüència, la probabilitat que a Madrid reben x_i trucades, sabent que a Barcelona es reben y_j trucades, és la probabilitat que a Madrid se'n reben x_i. Per tant, $P(X = x_i/Y = y_j) = P(X = x_i)$ per a qualsevol parell de valors (x_i, y_j). Aquesta expressió també es pot escriure segons la definició de *probabilitat condicionada* com: $P(X = x_i, Y = y_j) = P(X = x_i \cap Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$. Més formalment, si (X, Y) és una variable aleatòria bidimensional discreta, es diu que X i Y són independents si $P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$ per a qualsevol parell (x_i, y_j). Emprant notació amb funcions de probabilitat, la igualtat de probabilitats es tradueix en:

$$p(x_i, y_j) = p_x(x_i) \cdot p_y(y_j).$$

on p és la funció de probabiltat conjunta i p_x i p_y, les funcions de probabilitat marginals.

De manera anàloga, donada una variable aleatòria bidimensional contínua (X, Y), es diu que X i Y són independents si:

$$f(x, y) = f_X(x) \cdot f_Y(y).$$

On f és la funció de densitat conjunta i f_x i f_y, les funcions de densitat marginals.

Cal notar que aquestes definicions es poden generalitzar per a qualsevol nombre de variables aleatòries, així:

- Donada $(X_1,...,X_n)$ una variable aleatòria n-dimensional discreta, les variables aleatòries $X_1,...,X_n$ són independents si la funció de probabilitat conjunta és igual al producte de les funcions de probabilitat marginals:

$$p(x_1,...,x_n) = p_{X_1}(x_1) \cdot ... \cdot p_{X_n}(x_n).$$

- Anàlogament, donada $(X_1,...,X_n)$ una variable aleatòria n-dimensional contínua, les variables aleatòries $X_1,...,X_n$ són independents si la funció de densitat de probabilitat conjunta és igual al producte de les funcions de densitat de probabilitat marginals:

$$f(x_1,...,x_n) = f_{X_1}(x_1) \cdot ... \cdot f_{X_n}(x_n).$$

Nota

Cal dir que en les variables aleatòries bidimensionals es poden definir conceptes semblants als que es defineixen en les variables estadístiques bidimensionals, com són la covariància i la correlació lineal per a conèixer el grau de relació existent entre les variables. No obstant això, aquest estudi queda fora dels objectius d'aquest text.

Exemple 19

Un inversor té accions en dues empreses espanyoles. El rendiment de les accions, en percentatges, en cada empresa, són dues variables aleatòries X i Y. La distribució de probabilitat conjunta és la que apareix en la taula.

	Rendiment de *Y*			
Rendiment de *X*	5%	8%	12%	$P_x(x_i)$
5%	0,1	0,2	0,2	0,5
10%	0,04	0,08	0,08	0,2
12%	0,06	0,12	0,12	0,3
$P_y(y_j)$	0,2	0,4	0,4	1

Són les variables X i Y independents?

Les variables són independents si per a tots els possibles valors de la variable aleatòria bidimensional (x_i, y_j) es compleix que $p(x_i, y_j) = P_x(x_i)\, P_y(y_j)$:

- $p(5, 5) = 0{,}100$ i $P_x(5) \cdot P_y(5) = 0{,}50 \cdot 0{,}20 = 0{,}10 \rightarrow p(5{,}5) = P_x(5) \cdot P_y(5)$
- $p(5, 8) = 0{,}200$ i $P_x(5) \cdot P_y(8) = 0{,}50 \cdot 0{,}40 = 0{,}20 \rightarrow p(5{,}8) = P_x(5) \cdot P_y(8)$
- $p(5{,}12) = 0{,}2$ i $P_x(5) \cdot P_y(12) = 0{,}50 \cdot 0{,}40 = 0{,}20 \rightarrow p(5{,}12) = P_x(5) \cdot P_y(12)$
- $p(10{,}5) = 0{,}04$ i $P_x(10) \cdot P_y(5) = 0{,}2 \cdot 0{,}2 = 0{,}04 \rightarrow p(10{,}5) = P_x(10) \cdot P_y(5)$
- $p(10{,}8) = 0{,}08$ i $P_x(10) \cdot P_y(8) = 0{,}2 \cdot 0{,}4 = 0{,}08 \rightarrow p(10{,}8) = P_x(10) \cdot P_y(8)$
- $p(10{,}12) = 0{,}08$ i $P_x(10) \cdot P_y(12) = 0{,}2 \cdot 0{,}4 = 0{,}08 \rightarrow p(10{,}12) = P_x(10) \cdot P_y(12)$
- $p(12{,}5) = 0{,}06$ i $P_x(12) \cdot P_y(5) = 0{,}3 \cdot 0{,}2 = 0{,}06 \rightarrow p(12{,}5) = P_x(12) \cdot P_y(5)$
- $p(12{,}8) = 0{,}12$ i $P_x(12) \cdot P_y(8) = 0{,}3 \cdot 0{,}4 = 0{,}12 \rightarrow p(12{,}8) = P_x(12) \cdot P_y(8)$
- $p(12{,}12) = 0{,}12$ i $P_x(12) \cdot P_y(12) = 0{,}3 \cdot 0{,}4 = 0{,}12 \rightarrow p(12{,}12) = P_x(12) \cdot P_y(12)$.

En conseqüència, les variables *X* i *Y* són independents.

Exemple 20

Si *X* i *Y* representen la duració de dos determinats productes essencials per al funcionament d'una màquina i la funció de densitat conjunta és:

$$f(x, y) = e^{-(x+y)} \text{ on } x \geq 0 \text{ i } y \geq 0.$$

es pot dir que les variables són independents?

Com que $f(x, y) = e^{-(x+y)}$, es pot escriure en forma de producte de dues funcions, una que únicament depèn de *x* i l'altra de *y*: $f(x, y) = e^{-(x+y)} = e^{-x} e^{-y}$, llavors es compleix que *X* i *Y* són independents.

Nota

Encara que per a considerar que dues variables aleatòries són independents caldria ser exhaustiu i comprovar que formalment ho són –tal com s'ha fet en els exemples anteriors–, en moltes ocasions la independència pot considerar-se sense fer la comprovació, únicament per aplicació del sentit comú o per la informació de la qual es disposa.

7.4.4. Combinació lineal de variables aleatòries independents

En nombroses ocasions resulta molt interessant l'estudi de les funcions de variables aleatòries. Per exemple, si s'extrau a l'atzar una parella d'una gran població de parelles que treballen, per saber els ingressos conjunts de la parella. En aquest cas considerant les variables X = ingressos d'un membre de la parella i Y = ingressos de l'altre membre, els ingressos totals, que anomenarem S, és una funció de les dues variables, i es compleix que $S = X + Y$. O bé, si es vol saber la retenció de l'IRPF i se sap que a un dels membres, li retenen un 16% i a l'altre, un 17%, llavors la retenció serà de $W = 0{,}16 \cdot X + 0{,}18 \cdot Y$.

És evident que les variables unidimensionals resultants de les funcions de variables aleatòries també són variables aleatòries (tant S com W ho són). Conseqüentment, tenen una distribució de probabilitat que depèn de les variables de les quals són funcions. També els estadístics relacionats, com ara l'esperança i la variància de la nova variable, depenen de les esperances i les variàncies de les variables que l'originen. Així, en l'exemple, tant la distribució de probabilitat com l'esperança i la variància de S i W depenen de les distribucions i de les esperances i les variàncies de X i Y.

Aquestes qüestions poden estendre's per a més de dues variables. És a dir, donades $X_1,..., X_n$ n variables aleatòries i Z una funció d'aquestes ($Z = f(X_1,..., X_n)$), la distribució de probabilitat de Z depèn de les distribucions de probabilitat de cada variable $X_1,..., X_n$. Així mateix, l'esperança i la variància de Z depenen de les esperances i les variàncies de $X_1,...,X_n$. L'exemple següent aclareix aquests aspectes.

Nota

Encara que en el paràgraf anterior s'ha considerat una funció qualsevol, en el text únicament es tractaran funcions lineals, atès que aquest és un manual introductori. Cal recordar que les funcions lineals són aquelles en què les variables aleatòries no es multipliquen entre si, ni apareixen elevades a una potència. Si X_1, X_2 i X_3 són tres variables aleatòries, les variables $Z = 2 \cdot X_1 + X_2 + X_3$, $M = \frac{1}{2} X_1 - 3 \cdot X_2 + X_3 - 8$ són dos exemples de variables aleatòries construïdes com a funcions lineals de X_1, X_2 i X_3. També se sol dir que les variables Z i M són combinació lineal de les variables X_1, X_2 i X_3.

Exemple 21

Es consideren X i Y dues variables aleatòries que representen el nombre previst en grans inversions que tenen les dues empreses per a l'any vinent. La distribució de probabilitat de cadascuna és la següent:

X (primera empresa)	0	1	2	Y (segona empresa)	0	1	2	3
$p(x_i)$	0,2	0,5	0,3	$q(y_i)$	0,1	0,3	0,5	0,1

Totes dues empreses estan pensant de fusionar-se i per això volen saber com es distribueix el nombre total de grans projectes que durien a terme l'any següent. Per la informació que es maneja, ambdues variables poden considerar-se independents.

La variable aleatòria de la qual es demana la distribució de probabilitat és la suma de totes dues variables, X i Y, ja que aquestes són el nombre de grans projectes que té previst tenir cada empresa per a l'any següent. Es defineix, doncs, $S = X + Y$. A més a més, aquestes variables poden considerar-se independents, ja que totes dues empreses, en principi, no basen les seues decisions en el coneixement del que fa l'altra.

Llavors, el rang de S és $\{0, 1, 2, 3, 4, 5\}$ i la seua funció de probabilitat és:

- $p(0) = P(S = 0) = P(X = 0, Y = 0) = p(0)\, q(0) = 0{,}2 \cdot 0{,}1 = 0{,}02$.
- $p(1) = P(S = 1) = P(X = 0, Y = 1) + P(X = 1, Y = 0) = p(0)\, q(1) + p(1)\, q(0) = = 0{,}2 \cdot 0{,}3 + 0{,}5 \cdot 0{,}1 = 0{,}11$.
- $p(2) = P(S = 2) = P(X = 2, Y = 0) + P(X = 1, Y = 1) + P(\mathrm{X} = 0, Y = 2) = p(2)\, q(0) + p(1)\, q(1) + p(0)\, q(2) = 0{,}3 \cdot 0{,}1 + 0{,}5 \cdot 0{,}3 + 0{,}2 \cdot 0{,}5 = 0{,}28$.
- $p(3) = P(S = 3) = P(X = 2, Y = 1) + P(X = 1, Y = 2) + P(X = 0, Y = 3) = p(2)\, q(1) + p(1)\, q(2) + p(0)\, q(3) = 0{,}3 \cdot 0{,}3 + 0{,}5 \cdot 0{,}5 + 0{,}2 \cdot 0{,}1 = 0{,}36$.
- $p(4) = P(S = 4) = P(X = 1, Y = 3) + P(X = 2, Y = 2) = p(1)\, q(3) + p(2)\, q(2) = = 0{,}5 \cdot 0{,}1 + 0{,}3 \cdot 0{,}5 = 0{,}2$.
- $p(5) = P(S = 5) = P(X = 2, Y = 3) = p(2)\, q(3) = 0{,}3 \cdot 0{,}1 = 0{,}03$.

El que hom ara es pot preguntar és com es poden calcular la mitjana i la variància de variables construïdes com a combinació lineal d'altres. Cal dir que el càlcul d'aquests estadístics depèn del grau de relació existent entre les variables, en concret de si estan correlades o no. Això no obstant, com que es consideraran variables aleatòries independents, tant el càlcul com les propietats que tot seguit es mostren es redueixen notablement.

Propietats

Per a dues variables:

Es consideren dues variables aleatòries independents X_i i X_2. Es considera $Z = a \cdot X_2 + b \cdot X_2 + K$ on a, b i K són tres nombres distints de zero. Llavors:

1) $E(Z) = a \cdot E(X_1) + b\, E(X_2) + K$. En particular, $E(X_1 + X_2) = E(X_1) + E(X_2)$
2) $\text{Var}(Z) = a^2 \cdot \text{Var}(X_1) + b^2 \cdot \text{Var}(X_2)$. En particular, $\text{Var}(X_1 + X_1) = \text{Var}(X_1) + \text{Var}(X_2)$.

Per a més de dues variables:

Es consideren les variables aleatòries independents $X_1,...,X_n$, X_1 i X_2. Es considera $Z = a_1 \cdot X_1 + a_2 \cdot X_2 + a_nX_n + K$ on a_1, a_2,...,a_n i K són nombres distints de zero. Llavors:

1) $E(Z) = a_1 \cdot E(X_1) + a_2 \cdot E(X_2) + ... + a_nE(X_n) + K$
2) $\text{Var}(Z) = a^2_1 \cdot \text{Var}(X_1) + a^2_2 \cdot \text{Var}(X_2) + ... + a^2_n \text{Var}(X_n)$.

Exemple 22

Es consideren dues variables aleatòries X i Y, on X = els ingressos mensuals que cobra un dels membres d'una parella i Y = els ingressos de l'altre. Se sap que totes dues variables són independents i que el sou mitjà o esperat del primer és de 1.800 € i la variància, de 200 €^{2}. El valor esperat de l'altre és de 1.900 € i la variància és de 250 €^{2}. Quin és el sou esperat de la parella? Si el primer membre té una retenció de l'IRPF del 22% i el segon, del 21%, quants diners del sou retindran a la parella? Quines seran les variàncies?

En primer lloc, cal definir la variable «sou de la parella», el qual és evidentment $S = X + Y$. A més a més, com que les variables són independents, llavors aplicant les propietats s'obté que:

$$E(S) = E(X + Y) = E(X) + E(Y) \rightarrow \text{E}(S) = 1.800 + 1.900 = 3.600 \text{ €}$$
$$\text{Var}(S) = \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \rightarrow \text{Var}(S) = 200 + 250 = 450 \text{ €}^2.$$

Per a respondre a la segona pregunta cal definir la variable «retenció de la parella», la qual és $R = 0{,}22X + 0{,}21Y$. De la mateixa manera que en el cas antrior, aplicant-hi les propietats:

$$E(R) = E(0{,}22X + 0{,}21Y) = 0{,}22E(X) + 0{,}21E(Y)$$

$$E(S) = 0{,}22 \cdot 1.800 + 0{,}21 \cdot 1.900 = 795 \text{ €}$$

$$\text{Var}(R) = \text{Var}(0{,}22X + 0{,}21Y) = 0{,}22^2\text{Var}(X) + 0{,}212\text{Var}(Y)$$

$$\text{Var}(R) = 0{,}22^2 \cdot 200 + 0{,}21^2 \cdot 250 = 20{,}705\ €^2.$$

Exemple 23

Es consideren 30 variables aleatòries independents que representen el sou mensual de 30 persones respectivament. Si totes tenen la mateixa esperança (de 1.750 €) i la mateixa variància (de 230 €), quin serà el valor esperat de la mitjana de totes aquestes variables (la mitjana del sou de tots els treballadors) i quina la variància d'aquesta mitjana?

En primer lloc es defineix $\overline{X} = \frac{X_1 + X_2 + \ldots + X_{30}}{30} = \frac{X_1}{30} + \ldots + \frac{X_{30}}{30}$. Aplicant les propietats de l'esperança:

$$E\left(\overline{X}\right) = \left(\frac{X_1}{30} + \ldots + \frac{X_{30}}{30}\right) = \frac{1}{30}E(X_1) + \ldots + \frac{1}{30}E(X_{30}) = \frac{1}{30}30 \cdot (1.750) = 1.750\ €.$$

Pel que fa a la variància:

$$Var\left(\overline{X}\right) = Var\left(\frac{X_1}{30} + \ldots + \frac{X_{30}}{30}\right) = \left(\frac{1}{30}\right)^2 Var(X_1) + \ldots + \left(\frac{1}{30}\right)^2 Var(X_{30}) =$$

$$= \left(\frac{1}{30}\right)^2 30 \cdot (230) = 7{,}67\ €.$$

7.5 Models de probabilitat discrets: les distribucions de Bernoulli, binomial, hipergeomètrica i de Poisson

En els epígrafs anteriors s'han considerat les característiques més importants de les distribucions de probabilitat discretes des d'un punt de vista general. En aquest epígraf es tractaran alguns dels models de probabilitat discreta unidimensional més importants. Aquests són els models de Bernoulli, binomial, hipergeomètric i de Poisson.

7.5.1. La distribució de Bernoulli

La distribució de Bernoulli és una peça clau per a comprendre la distribució binomial, ja que aquesta última es pot considerar com una suma de distribucions de Bernoulli. Un experiment aleatori es distribueix segons un model de Bernoulli si l'experiment té únicament dos possibles resultats mútuament excloents. Aquests resultats se solen denominar *èxit* i *fracàs*. A més, si p és la probabilitat d'èxit, aleshores $1-p$ és la probabilitat de fracàs. Si ara es defineix la variable aleatòria X de manera que prenga el valor 1 si ix èxit i 0 si ix fracàs, llavors la distribució queda completament determinada.

Notació del model

Si es denota per X la variable aleatòria, X pot prendre els valors $\{1, 0\}$:

$$X \approx Be(p), \text{ on } p = \text{probabilitat d'èxit.}$$

Cal remarcar que p és el paràmetre de la distribució.

Funció de probabilitat d'una distribució de Bernoulli

Si se suposa que una variable aleatòria $X \approx Be(p)$, llavors, la funció de probabilitat és òbviament:

$$f(1) = p$$
$$f(0) = 1 - p.$$

Esperança i variància d'una distribució de Bernoulli

Utilitzant les definicions de *esperança* i *variància* de les distribucions discretes, junt amb la definició de la funció de probabilitat de la variable de Bernoulli, es poden calcular fàcilment l'esperança i la variància. Així si $X \approx Be(p)$, llavors:

$$\begin{aligned} &E(X) = p \\ &\text{Var}(X) = p(1-p). \end{aligned}$$

Exemple 24

Una venedora d'assegurances fa una venda el 70% de les vegades que ho intenta. Si es defineix la variable aleatòria X de manera que val 1 si es fa la venda i 0 si no es fa, calcula el valor esperat de X i la variància.

És obvi que $X \approx Be(0,7)$, per tant, aplicant el que hem calculat abans:

$$\begin{aligned} &E(X) = p = 0,7 \\ &\text{Var}(X) = p(1-p) = 0,7 \cdot 0,3 = 0,21 \end{aligned}$$

7.5.2. La distribució binomial

Una vegada explicats els conceptes més bàsics dels models de probabilitat de les distribucions discretes, així com el model de Bernoulli, cal presentar el model de probabilitat discret més important. Es tracta del model binomial. En aquest apartat es mostraran les característiques que han de complir els experiments per a poder modelar-se mitjançant aquest model. També es mostraran la funció de probabilitat del model i els estadístics més importants.

Un experiment es distribueix segons un model binomial si compleix les característiques següents:

- Consisteix a realitzar n repeticions independents d'una experiència de Bernoulli.
- Aquest segon experiment, com s'ha explicat anteriorment, té dos possibles resultats (éxit/fracàs), i en cada repetició la probabilitat d'èxit (p) no varia, és constant. (Les experiències de Bernoulli són independents.)
- L'objectiu de l'experiment és comptar el nombre d'èxits obtinguts en les n repeticions.

Exemple 25

L'experiment «llançar 30 vegades una moneda equilibrada i comptar el nombre de cares que s'obtenen», és un experiment que segueix un model binomial, perquè:

- Es realitza 30 vegades un experiment que té dos possibles resultats: cara i creu.
- En cada realització de l'experiment, la probabilitat que isca cara $\left(\frac{1}{2}\right)$ no canvia.
- L'objectiu és comptar el nombre de cares.

En aquestes condicions, el model de probabilitat permet calcular les probabilitats d'obtenir un nombre determinat de cares en les 30 realitzacions.

Notació del model

Si es denota per X la variable aleatòria:

X = nombre d'èxits obtinguts en els n llançaments, llavors es representa per:

$$X \approx Bi(n, \text{p}) \text{ on } n = \text{nombre de realitzacions}$$
$$p = \text{probabilitat d'èxit}$$

Funció de probabilitat d'una distribució binomial

Si se suposa que una variable aleatòria $X \approx Bi(n, \text{p})$, llavors X pot prendre els valors $\{0, 1, 2, ..., n\}$, ja que X és el nombre d'èxits varia entre 0 i n.

Es pot demostrar matemàticament que la funció de probabilitat és:

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{si} \quad x = 0, 1, 2, 3, ...n \\ 0 & \text{la resta} \end{cases}$$

La representació gràfica, com ja s'havia comentat, és un diagrama de barres (figura 16).

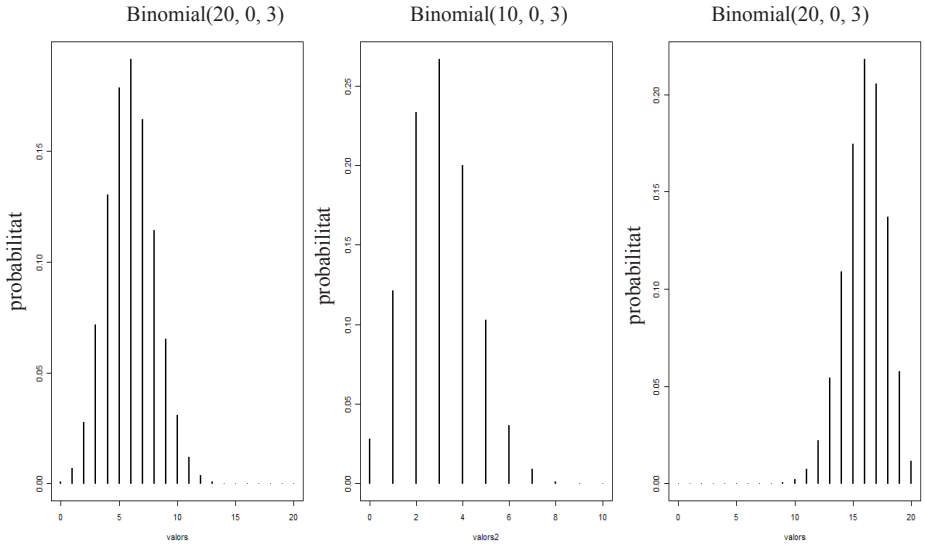


Figura 16

Exemple 26

L'experiment esmentat en l'exemple 25, $X \approx Bi(30,0{,}5)$. Per tant, la funció de probabilitat de la variable X és:

$$f(x) = \begin{cases} \binom{30}{x} 0{,}5^x (1-0{,}5)^{n-x} & \text{si} \quad x = 0, 1, 2, 3, ...n \\ 0 & \text{la resta} \end{cases}$$

D'aquesta manera, és molt senzill calcular la probabilitat d'obtenir 10 cares, simplement cal substituir en la fórmula x per 10.

$$P(X = 10) = \binom{30}{10} 0{,}5^{10} (1-0{,}5)^{20} = \frac{30!}{10! \cdot 20!} 0{,}5^{10} (1-0{,}5)^{20} = 0{,}02798$$

Esperança i variància d'una distribució binomial

Utilitzant les definicions de *esperança* i *variància* de les distribucions discretes, junt amb la definició de la funció de probabilitat de la variable binomial, es pot demostrar matemàticament que l'esperança i la variància són:

$$E(X) = np$$
$$\text{Var}(X) = np(1 - p).$$

Exemple 27

En l'experiment de l'exemple 25. $\approx Bi(30,0{,}5)$, es té que:

$$E(X) = np = 30 \cdot 0{,}5 = 15$$
$$\text{Var}(X) = np(1 - p) = 30\ 0{,}5 \cdot (1 - 0{,}5) = 7{,}5.$$

*Exemple 2*8

En una població, el 35% són simpatitzants del partit polític A. S'hi seleccionen 15 persones a l'atzar.

- Quina és la probabilitat que entre elles hi haja 10 simpatitzants del partit A?
- I que n'hi haja menys de 6?
- Quin és el nombre esperat de simpatitzants del partit 1 entre els 15 escollits?

Es defineix X = nre. de simpatitzants del partit A.

La probabilitat d'escollir-hi un simpatitzant del partit A és de 0,35 (el 35% ho són). Es pretén comptar el nombre de simpatitzants que hi ha entre els 15 i la probabilitat de ser simpatitzant del partit A no varia en cada una de les 15 seleccions. La raó és que es considera una població molt gran, i aleshores la probabilitat que la primera persona seleccionada siga simpatitzant del partit A (primer experiment) no es pot considerar que influïsca en la probabilitat que la següent persona també ho siga. Per exemple, si la població tinguera 1.000.000 d'habitants, 350.000 serien simpatitzants del partit A. Llavors les probabilitats condicionades de la segona selecció serien:

$$P(\text{2n afí partit/1r afí partit}) = \frac{349.999}{999.999} \cdot = 0{,}34999935$$

$$P(\text{2n afí partit/1r no afí partit}) = \cdot \frac{350.000}{999.999} = 0{,}35000035.$$

És a dir, en ambdós casos la diferència respecte a 0,35 és pràcticament menyspreable. Per a la resta de seleccions les probabilitats condicionades donen valors molt semblants a 0,35. Per tant, es pot concloure que totes les seleccions són independents i, en conseqüència, es compleixen tots els supòsits de la distribució binomial $X \approx Bi(15,0.35)$.

Així doncs, cal emprar la funció de probabilitat:

- La primera pregunta demana $P(X = 10) = \binom{15}{10} (0{,}35)^{10} \cdot (0{,}65)^5 = 0{,}00962$.
- La segona pregunta demana:

  $P(X < 6) = P(X = 0) + P(X = 1) + P(X = 2) + (X = 3) + P(X = 4) + P(X = 5)$

  $$= \sum_{k=0}^{5} \binom{15}{k} (0.35)^k (0.65)^{15-k} =$$

  $$= \binom{15}{0} 0{,}35^0 0{,}65^{15} + \binom{15}{1} 0{,}35^1 0{,}65^{14} + \binom{15}{2} 0{,}35^2 0{,}65^{13} + \binom{15}{3} 0{,}35^3 0{,}65^{12}$$

  $$+ \binom{15}{4} 0{,}35^4 0{,}65^{11} + \binom{15}{5} 0{,}35^5 0{,}65^{10} = 0{,}5643.$$

- La tercera pregunta demana $E(X) = n \cdot p = 15 \cdot 0{,}35 = 5{,}25$ persones.

Nota

Si en l'exemple 17 un dia se seleccionen 15 participants i al dia següent se'n seleccionen 25 més i es vol calcular la probabilitat de trobar 19 simpatitzants del partit A entre tots dos dies, com es pot fer? És a dir, com es distribueix el nou experiment format per la suma de dues variables aleatòries binomials, una $X_1 \approx Bi(15, 0{,}35)$ i l'altra $X_2 \approx Bi(10, 0{,}35)$?

La resposta en aquest cas és la que la intuïció pareix indicar. És a dir, el nombre de simpatitzants totals, $X_1 + X_2 \approx Bi(40, 0{,}35)$. No obstant això, cal tenir en compte dues coses: en primer lloc, totes dues variables aleatòries són independents, és a dir, el resultat de l'una no influeix en el de l'altra ($P(X_1 = a, X_2 = b) = P(X_1 = a) \cdot P(X_2 = b)$ per a qualssevol valors *a* i *b*). Si aquest fet no es produïra, no es compliria un dels supòsits de la distribució binomial per a $X_1 + X_2$, ja que totes les experiències de Bernoulli de $X_1 + X_2$ no tindrien la mateixa probabilitat d'èxit. En segon lloc, totes dues distribucions tenen la mateixa probabilitat d'èxit. En aquest cas, 0,35. Si aquest fet no es complira, tampoc en $X_1 + X_2$ les experiències de Bernoulli tindrien les mateixes probabilitats.

Aquestes dues exigències són les que permeten que $X_1 + X_2 \approx Bi(40, 0{,}35)$. A més a més, la propietat que s'acaba de comentar pot generalitzar-se per a qualsevol parell de distribucions binomials que complisquen ambdues exigències. És a dir:

$$\left.\begin{array}{l} Si\ X_1 \approx Bi(n_1, p)\ i \\ \quad X_2 \approx Bi(n_2, p) \\ \\ \quad X_1\ i\ X_2\ són \\ \quad independents \end{array}\right\} \Rightarrow X_1 + X_2 \approx Bi\,(n_1 + n_2, p)$$

Així doncs, responent a la pregunta inicial,

$$P(X_1 + X_2 = 19) = \binom{40}{19} (0.35)^{19} \cdot (0.65)^{21} = 0{,}0336$$

Cal dir que aquesta propietat es pot generalitzar per a més de dues variables aleatòries sempre que es complisquen ambdues exigències per a totes les variables.

7.5.3. La distribució hipergeomètrica

Un dels supòsits de la distribució binomial és que la probabilitat d'èxit en cada experiència de Bernoulli és constant. Si aquest supòsit no s'exigeix, llavors no es pot considerar una distribució binomial. Si aquest és l'únic supòsit que deixa de complir-se, aleshores la distribució de probabilitats s'anomena *hipergeomètrica*.

En l'exemple 28, s'ha raonat que en una població molt gran la probabilitat d'obtenir èxit en una de les eleccions no influeix en la probabilitat d'obtenir-ne en les següents. Si ara es consideren 11 persones on 7 són simpatitzants del partit A i la resta no, i es vol saber la probabilitat d'obtenir 4 simpatitzants en extraure 5 persones d'aquest grup, es pot considerar un model binomial? L'únic supòsit que pot no complir-se és el de la independència de les seleccions. És a dir, es

poden considerar les probabilitats condicionades als resultats anteriors constants? La resposta és que no:

$$P(\text{2n afí partit/1r afí partit}) = \frac{6}{10} = 0{,}6$$

$$P(\text{2n afí partit/1r no afí partit}) = \frac{7}{10} = 0{,}7.$$

Ambdues probabilitats són considerablement diferents de la probabilitat d'èxit (la persona escollida és afí al partit A) en la primera selecció $\frac{7}{11} = 0{,}636363...$ És evident, doncs, que en aquest tipus d'experiments no es pot emprar la distribució binomial. Però com es pot calcular la probabilitat del que se demana en aquest experiment? La manera més natural és emprant Laplace:

Casos possibles: $\binom{11}{5}$. Són tots els grups de 5 persones que poden fer-se amb 11.

Casos favorables: $\binom{7}{4}\cdot\binom{4}{1}$. Són tots els grups de 4 persones que poden fer-se amb 7 membres de manera que totes quatre siguen afins al partit A multiplicat per tots els grups d'una persona (la que falta per a tenir un grup de 5) que no siguen afins al partit A.

I, per tant: $P(\text{obtenir 4 simpatitzants}) = \dfrac{\binom{7}{4}\cdot\binom{4}{1}}{\binom{11}{5}} = 0{,}30303.$

Aquesta és la idea en la qual se sustenta el model hipergeomètric.

Notació del model

Si es denota per X la variable aleatòria, X = nombre d'èxits, llavors es representa per:

$X \approx HG(n, m, r)$, on
n = nombre total d'elements
m = nombre d'elements del subconjunt
r = nombre d'elements del total n que són èxits o que compleixen la condició.

Funció de probabilitat d'una distribució hipergeomètrica

Si se suposa que una variable aleatòria $X \approx HG(n, m, r)$, llavors X pot prendre els valors $\{0, 1, 2,..., r\}$, ja que X (nombre d'èxits) varia entre 0 i r.

Es pot demostrar matemàticament que la funció de probabilitat és:

$$P(X=k)=f(x)=\begin{cases}\dfrac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}} & \text{per a } k=1,...,r \text{ i } r \geq k.\\ 0 & \text{per a la resta.}\end{cases}$$

La representació gràfica és un diagrama de barres (figura 17).

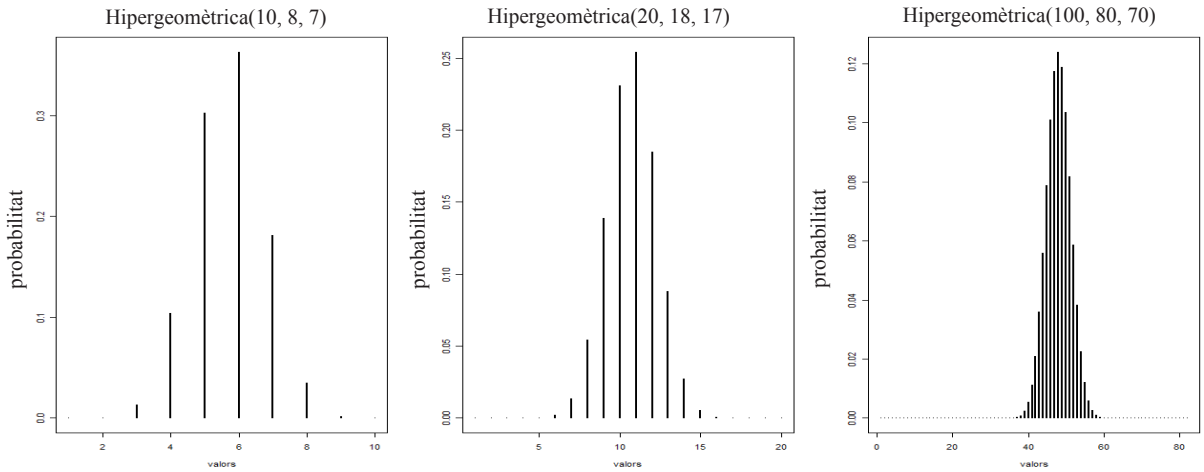


Figura 17

Exemple 29

En una empresa se seleccionaran a l'atzar 6 treballadors d'un total de 20 que té la plantilla, per a fer una anàlisi sobre el grau d'estudis que tenen els treballadors de les empreses de la localitat. Si dels 20 treballadors de l'empresa n'hi ha 8 que tenen estudis superiors, quina és la probabilitat que n'hi hagen 4 entre els 6 seleccionats?

Si es defineix X = nombre de treballadors amb estudis superiors, i es considera «èxit» = el treballador seleccionat té estudis superiors; l'experiment consisteix a repetir 6 vegades un model de Bernoulli: saber si el treballador té estudis superiors o no. A més a més, és clar que el resultat de cada experiment influeix en el resultat del següent, ja que hi ha un total de 20 persones.

En conseqüència, la variable aleatòria X es distribueix segons un model hipergeomètric. En concret, $X \approx HG(20,6,8)$.

$$\text{Per tant, } P(X=4) = f(4) = \frac{\binom{8}{4}\binom{12}{2}}{\binom{20}{6}} = 0{,}119.$$

Esperança i variància d'una distribució hipergeomètrica

Utilitzant les definicions de *esperança* i *variància* de les distribucions discretes, junt amb la definició de la funció de probabilitat de la variable hipergeomètrica, es pot demostrar matemàticament que l'esperança i la variància són:

$$E(X) = \frac{m \cdot r}{n}$$

$$\text{Var}(X) = \frac{n-m}{n-1} \cdot \frac{m \cdot r}{n} \cdot \left(1 - \frac{r}{n}\right).$$

Exemple 30

En l'exemple 29, es poden calcular l'esperança i la variància. Així, com que $X \approx HG(20,6,8)$, s'obté que:

$$E(X) = \frac{6 \cdot 8}{20} = 2{,}4 \text{ treballadors amb estudis superiors}$$

$$\text{Var}(X) = \frac{20-6}{19} \cdot \frac{6 \cdot 8}{20} \cdot \left(1 - \frac{8}{20}\right) = 1{,}061 \text{ treballadors}^2.$$

Exemple 31

En una fase d'una oposició, els aspirants han de desenvolupar un tema. Els opositors han d'escollir-ne un dels cinc que el tribunal tria d'una manera aleatòria d'un conjunt de 72 temes. Si una persona es presenta a l'oposició i se n'ha estudiat 30, calcula quina és la probabilitat que s'haja estudiat almenys un dels cinc que ha extret el tribunal. Calcula també el nombre esperat de temes que li han «encertat», a l'aspirant, dels cinc triats pel tribunal a l'atzar.

L'objectiu és comptar el nombre de temes que li són encertats (èxits) a l'aspirant. De l'enunciat es dedueix que dels 72 temes, 30 compleixen la condició «han estat estudiats per l'aspirant». A més a més, s'extrauen aleatòriament cinc temes dels 72. Consegüentment, si X = nombre de temes encertats a l'aspirant, $X \approx HG(72,5,30)$.

La primera pregunta demana:

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - \frac{\binom{30}{0}\binom{42}{5}}{\binom{72}{5}} = 1 - 0{,}061 = 0{,}939.$$

La segona pregunta demana el nombre esperat d'encerts. Aplicant-hi el que hem ressenyat amb anterioritat respecte a l'esperança d'una distribució hipergeomètrica:

$$E(X) = \frac{5 \cdot 30}{72} = 2{,}08 \text{ temes encertats.}$$

7.5.4. La distribució de Poisson

La distribució de Poisson fou proposta per primera vegada per Siméon Poisson (1781-1840) en un llibre publicat en 1837. La distribució de Poisson és una important distribució de probabilitat discreta per a nombroses aplicacions, com poden ser el nombre d'errades d'un gran sistema informàtic, el nombre de vaixells que arriben a un port en un període de 6 hores, el nombre de camions de repartiment que arriben a un magatzem en una hora, el nombre de defectes d'una peça metàl·lica, el nombre de clients que arriben a una caixa en un supermercat en un interval concret de temps, etc.

Així, es pot emprar la distribució de Poisson en les distribucions que es caracteritzen per ser el nombre d'ocurrències o èxits d'un esdeveniment en un interval continu donat (com ara el temps, la superfície o la longitud). En resum, la distribució de Poisson mesura la probabilitat d'un esdeveniment aleatori sobre algun interval continu.

Són necessaris tres supòsits per a poder aplicar la distribució de Poisson. En primer lloc, se suposa que l' interval està dividit en una gran quantitat de subintervals de la mateixa amplària, en els quals la probabilitat que es done una ocurrència és ínfima. Tenint en compte aquesta consideració els supòsits són:

- La probabilitat que ocórrega un esdeveniment és constant en tots els subintervals.
- No pot haver-hi més d'una ocurrència en cada subinterval.
- Les ocurrències són independents, és a dir, les ocurrències en intervals que no se solapen son independents entre si.

Notació del model

Si es denota per X la variable aleatòria,

X = nombre d'ocurrències en un interval, llavors es representa per:

$X \approx Po(\mu)$, on

μ = mitjana d'esdeveniments per unitat de temps o espai.

Funció de probabilitat d'una distribució de Poisson

Si es denota per X la variable aleatòria, llavors X pot prendre els valors {0, 1, 2, 3, 4...} i la funció de probabilitat de la distribució de Poisson es representa mitjançant un diagrama de barres (figura 18) i s'expressa com:

$$P(X = x) = f(x) = \frac{\mu^x \cdot e^{-\mu}}{x!}.$$

On: x és el nombre de vegades que ocorre l'esdeveniment.

μ mitjana d'esdeveniments per unitat de temps o espai.

$X \approx Po(\mu)$.

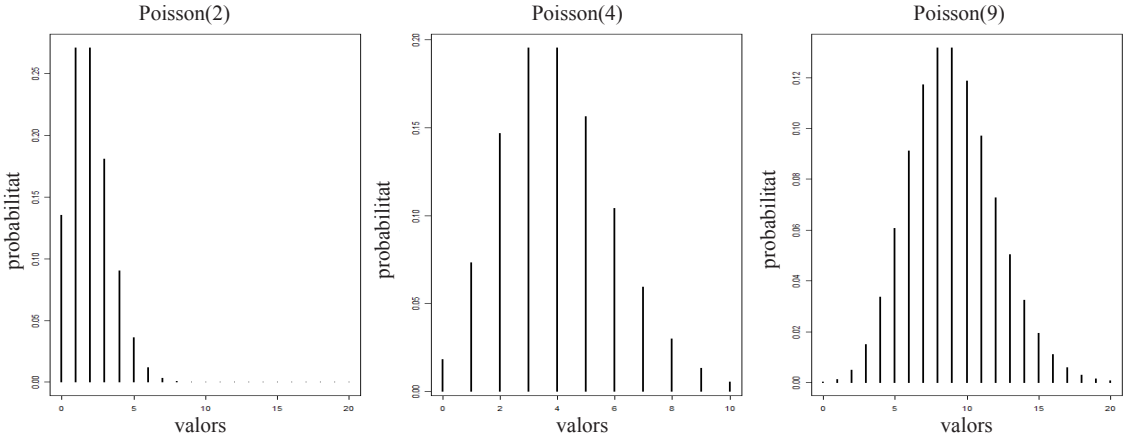


Figura 18

Exemple 32

En una tenda s'ha fet un estudi i s'ha recollit que al llarg de les darreres 80 hores hi han entrat 800 clients. Calcula la probabilitat que durant l'hora següent entren cinc clients.

Si X = nombre de clients que entren en una hora a l'establiment, i hi han entrat 800 clients en una hora (la mitjana de clients per hora és de deu clients) $X \approx Po(10)$.

Per tant, emprant-hi la fórmula: $P(X = 5) = f(5) = \frac{10^5 \cdot e^{-10}}{5!} = 0{,}0378$.

Cal notar que aquests càlculs es poden realitzar directament o mitjançant programes informàtics de caire estadístic. També algunes calculadores incorporen els procediments estadístics adequats. Per a acabar, també es poden fer els càlculs mitjançant taules estadístiques.

Esperança i variància d'una distribució de Poisson

Utilitzant les definicions de *esperança* i *variància* de les distribucions discretes, junt amb la definició de la funció de probabilitat de la variable de Poisson, es pot demostrar matemàticament que l'esperança i la variància són:

$$E(X) = \mu$$
$$Var(X) = \mu.$$

Exemple 33

Els clients d'un bar arriben a una màquina de cafè de manera que aquesta ha de traure una mitjana de dos cafès cada cinc minuts. Calcula la probabilitat que la màquina haja de traure més de dos cafès en un període de 5 minuts.

Es defineix X = nombre de cafès que serveix la màquina durant cinc minuts. Considerant que es compleixen els supòsits, $X \approx Po(2)$. Es demana $P(X > 2)$. Per a calcular-ho, s'aplica la regla del complementari, és a dir:

$$P(X > 2) = 1 - P(X \leq 2) = 1 - (P(X = 0) + P(X = 1) + P(X = 2)) =$$

$$= 1 - \left(\frac{2^0 \cdot e^{-2}}{0!} + \frac{2^1 \cdot e^{-2}}{1!} + \frac{2^2 \cdot e^{-2}}{2!} \right) =$$

$$= 1 - (0{,}1353 + 0{,}2707 + 0{,}2707) = 0{,}3233.$$

Aproximació de la binomial per la distribució de Poisson

La distribució de probabilitats de Poisson està molt relacionada amb la distribució binomial. De fet, quan el nombre de repeticions de la experiencia de Bernoulli és molt elevat i la probabilitat d'èxit, reduïda, es pot aproximar una distribució per l'altra.

Així doncs, en els casos en què el nombre de repeticions n i la probabilitat d'èxit p complisquen que $np \leq 7$ pot aproximar-se una distribució $Bi(n, p)$ per una distribució $Po(np)$. És a dir, per una distribució de Poisson de paràmetre np. La figura 19 reflecteix aquesta relació.

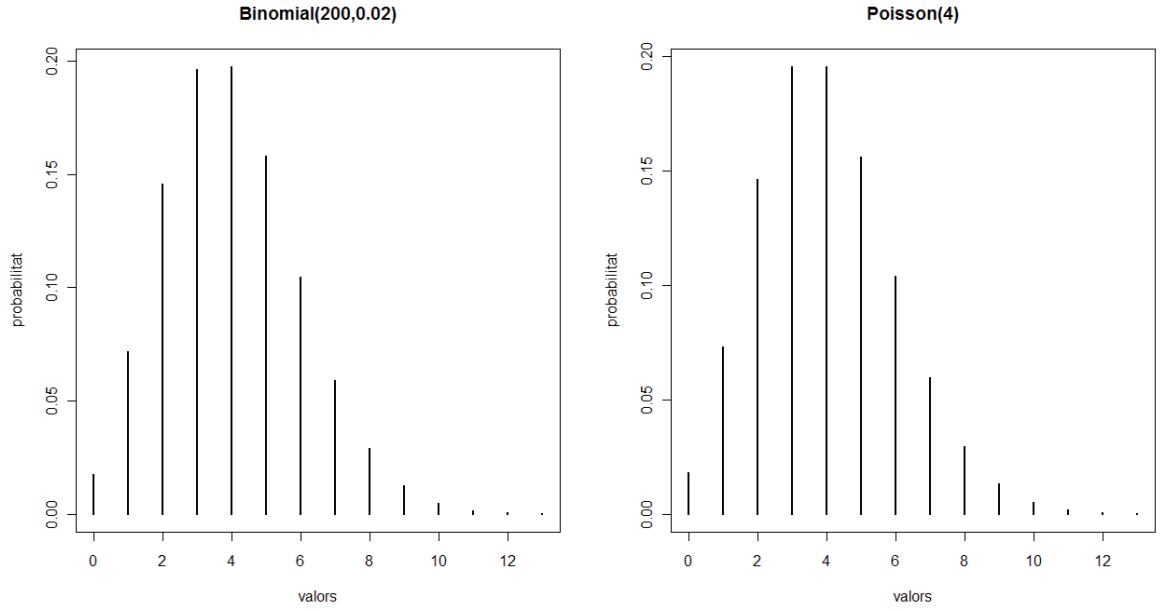


Figura 19

Exemple 34

Un analista ha pronosticat que el 3,5% de les empreses d'un país reduirà la seua plantilla enguany. Suposant que la previsió siga correcta, calcula la probabilitat que de 100 empreses del país preses a l'atzar, almenys tres reduïsquen la plantilla.

S'assigna X = nombre d'empreses que reduiran la plantilla.

És clar que $X \approx Bi(100, 0{,}035)$, ja que l'experiment consisteix a repetir 100 vegades l'experiment d'observar si una empresa reduirà la plantilla (èxit) o no (fracàs). La probabilitat d'èxit és de 0,035 i totes les repeticions de l'experiment són independents les unes de les altres.

Per altra part, com que $n \cdot p = 100 \cdot 0{,}035 = 3{,}5 < 7$, llavors la distribució binomial pot aproximar-se per una de Poisson de paràmetre 3,5 ($X \approx Po(3{,}5)$).

Així doncs, $P(X \geq 3) = 1 - P(X \leq 2) = 1 - (P(X = 0) + P(X = 1) + P(X = 2))$.

I com que $P(X = 0) = \dfrac{e^{-3{,}5}\,3{,}5^0}{0!} = e^{-3{,}5} = 0{,}0302$

$$P(X = 1) = \frac{e^{-3{,}5}\,3{,}5^1}{1!} = 3{,}5e^{-3{,}5} = 0{,}1057$$

$$P(X = 2) = \frac{e^{-3{,}5}\,3{,}5^2}{2!} = 0{,}1850,$$

llavors $P(X \geq 3) = 1 - (0{,}0302 + 0{,}1057 + 0{,}1850) = 0{,}67916$.

Cal remarcar que aplicant-hi la distribució binomial, $P(X \geq 3) = 0{,}684093$. Per tant, com es pot comprovar, l'aproximació és bastant bona.

Suma de variables de Poisson

La distribució de Poisson compleix la mateixa propietat que les distribucions binomials pel que fa a la suma de distribucions de Poisson independents, és a dir:

$$\left.\begin{array}{l}\text{Si } X_1 \approx Po(\lambda_1) \text{ i} \\ \quad X_2 \approx Po(\lambda_2) \\ \\ X_1 \text{ i } X_2 \text{ són} \\ \text{independents}\end{array}\right\} \Rightarrow X_1 + X_2 \approx Po(\lambda_1 + \lambda_2)$$

Així, per exemple, si el nombre de trucades de telèfon en una hora segueix una distribució de Poisson de mitjana 12 i el nombre de trucades en la mateixa hora del dia següent segueix una distribució de Poisson de vuit telefonades de mitjana, llavors el nombre de trucades entre les dues hores segueix una distribució de Poisson de mitjana 20. Formalment, $X_1 \approx Po(12)$ i $X_2 \approx Po(8)$ $X_1 + X_2 \approx Po(20)$.

7.6. Models de probabilitat continus: les distribucions uniforme i exponencial

En els epígrafs anteriors s'han considerat les característiques més importants de les distribucions de probabilitat discreta des d'un punt de vista general. En aquest epígraf es tractaran alguns dels models de probabilitat contínua unidimensional més importants. Aquests són els models uniforme i exponencial. El model normal es tracta en un epígraf específic per la seua importància.

7.6.1. La distribució uniforme

L'exemple 13 pot generalitzar-se per a qualsevol experiment en què la variable siga contínua i, els valors que hi pot prendre siguen aleatoris dins d'un interval. Així, una variable aleatòria X segueix una distribució uniforme a l'interval $[a, b]$ i es representa per $U[a, b]$, si és contínua i tots els valors de l'interval tenen les mateixes possibilitats de ser presos per la variable aleatòria.

Notació del model

Si es denota per X la variable aleatòria:

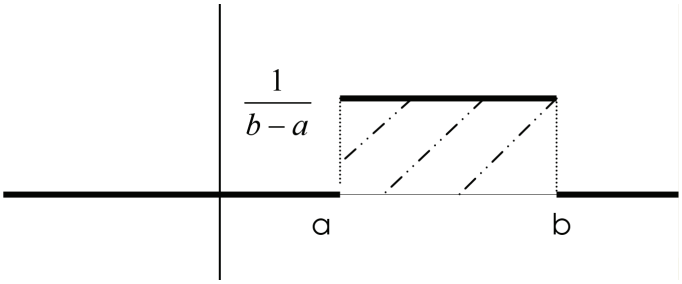
X = valor escollit aleatòriament de l'interval $[a, b]$, es representa per $X \approx U[a, b]$.

Funció de densitat d'una distribució $U[a, b]$

La funció de densitat és:

$$f(t) = \begin{cases} \dfrac{1}{b-a} & \text{si} \quad a \le t \le b \\ 0 & \text{si} \quad t < a \ \ o \ \ t > b. \end{cases}$$

I la representació gràfica és:



A més a més, donats x_1 i x_2 dos valors de l'interval $[a, b]$, es pot demostrar mitjançant càlcul integral que:

$$\mathrm{P}(x_1 \le X \le x_2) = \frac{x_2 - x_1}{b-a}.$$

En conseqüència, $P(X \le x_1) = P(a \le X \le x_1) = \dfrac{x_1 - a}{b-a}$.

Esperança i variància d'una distribució $U[a, b]$

Es pot demostrar que, si una variable aleatòria es distribueix uniformement en l'interval $[a, b]$, llavors:

$$E(X) = \mu = \frac{a+b}{2}$$

$$Var(X) = \sigma^2 = \frac{(b-a)^2}{12}.$$

Exemple 35

Se sap que el conductor d'un autobús tarda un mínim de 63 minuts i un màxim de 110 minuts a realitzar un trajecte. Si es coneix que el temps de conducció està distribuït uniformement, calcula el temps mitjà que empra el conductor a realitzar el trajecte i la probabilitat que el temps del trajecte siga inferior a 90 minuts.

Es defineix X = temps que li consta, al conductor, realitzar el trajecte. L'enunciat diu que X es distribueix segons $U(63, 110)$.

La primera pregunta demana $E(X) = \frac{63 + 110}{2} = 86{,}5$ minuts.

La segona pregunta demana $P(X \leq 90) = \frac{90 - 63}{110 - 63} = \frac{17}{37} = 0{,}459$.

7.6.2. La distribució exponencial

La distribució exponencial se sol emprar per a resoldre problemes de llistes d'espera o cues. D'alguna manera, es pot considerar com a complementària de la distribució de Poisson, ja que si aquesta calcula la probabilitat que ocórreguen un nombre d'esdeveniments en un interval de temps o espai en el qual se'n produeixen una mitjana de μ, la distribució exponencial calcula la probabilitat del temps que ha de transcórrer entre dues ocurrències si el nombre d'ocurrències per unitat de temps és μ.

Notació del model

Si es denota per X la variable aleatòria:

X= temps que passa entre dues ocurrències, es representa per:

$$X \approx \text{Exp}(\mu),$$

on μ = mitjana d'esdeveniments per unitat de temps o espai.

Funció de densitat d'una distribució Exp(μ)

La funció de densitat és:

$$f(t) = \begin{cases} \mu \cdot e^{-\mu t} & t \geq 0 \\ 0 & t < 0 \end{cases},$$

on t és el nombre d'unitats de temps fins a l'ocurrència següent,

i μ és el nombre mitjà d'ocurrències per unitat de temps.

Es pot demostrar matemàticament que la funció de distribució és $F(t)=1-\cdot e^{-\mu t}$. La representació gràfica d'ambdues funcions és la següent:

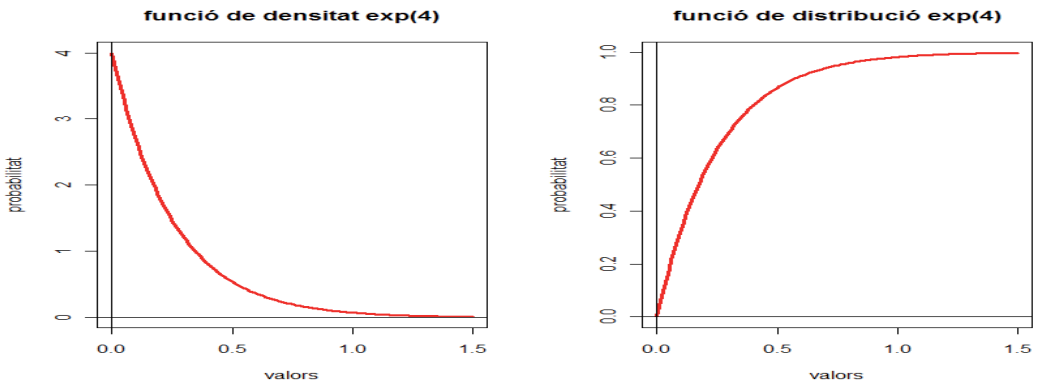


Figura 20

Esperança i variància d'una distribució Exp(μ)

Es pot demostrar que si una variable aleatòria es distribueix exponencialment de paràmetre μ llavors:

$$E(X) = \frac{1}{\mu}$$

$$Var(X) = \sigma^2 = \frac{1}{\mu^2}.$$

Exemple 36

El temps que dedica un metge a atendre els pacients pot representar-se per una distribució exponencial que té un temps mitjà d'atenció de 5 minuts. Quina serà la probabilitat que el temps d'atenció siga superior a 10 minuts.

Si t és el temps d'atenció en minuts, el nombre d'atencions per minut és $\frac{1}{5} = 0{,}2$ i la funció de densitat és $f(t) = 0{,}2 \cdot e^{-0{,}2t}$.

Llavors, la probabilitat que el temps d'atenció siga superior a 10 minuts es calcula així:

$$P(X > 10) = 1 - P(X < 10) = 1 - F(10) = 1 - (1 - 0{,}2 \cdot e^{-0{,}2 \cdot 10}) = 0.1353.$$

7.7. Distribució normal. El teorema del límit central

La distribució normal és, sens dubte, la més important de totes les distribucions de probabilitat del càlcul de probabilitats i de l'estadística. Hi ha tres raons principals:

- Les propietats matemàtiques. És d'enorme importància en la inferència estadística.
- L'aplicació. Un gran nombre de fenòmens reals es poden modelar mitjançant la distribució normal. Per exemple, la distribució d'alçades, pesos, distàncies i altres variables que són divisibles infinitament.
- El teorema del límit central. La distribució normal serveix per a aproximar la suma i la mitjana de qualsevol altre tipus de distribucions.

Cal dir que la distribució normal necessita dos paràmetres per a poder ser definida. Així, és necessari conèixer la mitjana o esperança i la desviació típica per a poder calcular el model. Aquest fet és en moltes ocasions impossible, i s'hi fa necessari l'ús d'estimacions puntuals d'aquests paràmetres o d'altres tècniques d'inferència. No obstant això, com que es definirà la distribució normal des d'un punt teòric, es consideraran coneguts tots dos paràmetres.

7.7.1. Definició

Una variable aleatòria X segueix una distribució normal amb paràmetres μ i σ ($-\infty < \mu < +\infty$ *i* $\sigma > 0$), i es denota per $X \sim N(\mu, \sigma)$, si la funció de densitat de probabilitat és:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]} \qquad \text{per a } -\infty < x < \infty$$

Per ser contínua, la probabilitat que la variables aleatòria X estiga compresa entre dos valors x_1 i x_2, es calcula mitjançant el càlcul integral:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{(t-\mu)^2}{2\sigma^2}\right]} dt.$$

No obstant això, existeix un procediment que es desenvoluparà més endavant, que permet el càlcul d'aquestes probabilitats emprant una taula i, en conseqüència, sense resoldre la integral. La representació gràfica de la funció de densitat té forma de campana. La figura 21 és la funció de densitat d'una distribució normal de paràmetres (10,5).

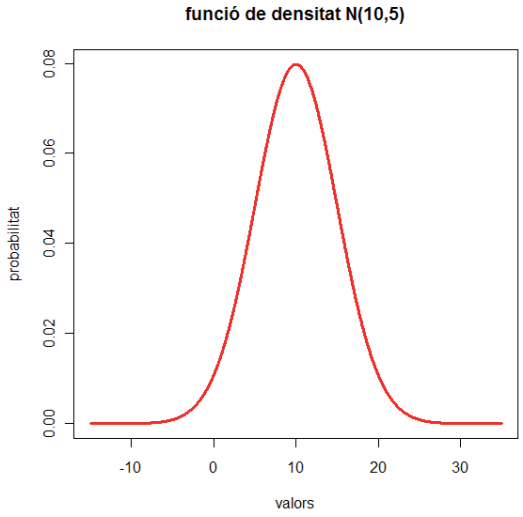


Figura 21

Esperança i variància d'una distribució normal

Mitjançant càlculs matemàtics, es pot demostrar que:

$E(X) = \mu$

$Var(X) = \sigma^2$ i llavors la desviació típica de la distribució és σ.

És a dir, els paràmetres que defineixen una distribució de probabilitat normal són l'esperança (μ) i la desviació típica (σ).

Exemple 37

El temps que tarda un ordinador d'una determinada marca a espatllar-se segueix una distribució normal amb una esperança de 410 dies i una desviació típica de 50 dies. Calcula la probabilitat que l'ordinador s'espatlle abans que finalitze un any.

Es defineix X = temps que tarda un ordinador a espatllar-se. Per tant, $X \sim N(410, 50)$.

El problema demana $P(X \leq 365)$. Per a calcular-ho, amb la informació proporcionada fins ara, cal realitzar el càlcul integral així:

$$P(X \leq 365) = \int_{-\infty_1}^{365_2} \frac{1}{50\sqrt{2\pi}} e^{\left[-\frac{(t-410)^2}{2\cdot 50^2}\right]} dt = 0{,}1841.$$

Nota

Cal notar que per a poder realitzar els càlculs de probabilitat emprant les taules de la distribució normal o el software informàtic adient, és convenient conèixer la gràfica de la funció de densitat de probabilitat, així com algunes de les propietats més característiques.

Representació gràfica. Interpretació dels paràmetres

Es mostra tot seguit la representació gràfica de la funció de densitat d'una distribució normal d'esperança 0 i desviació típica 1.

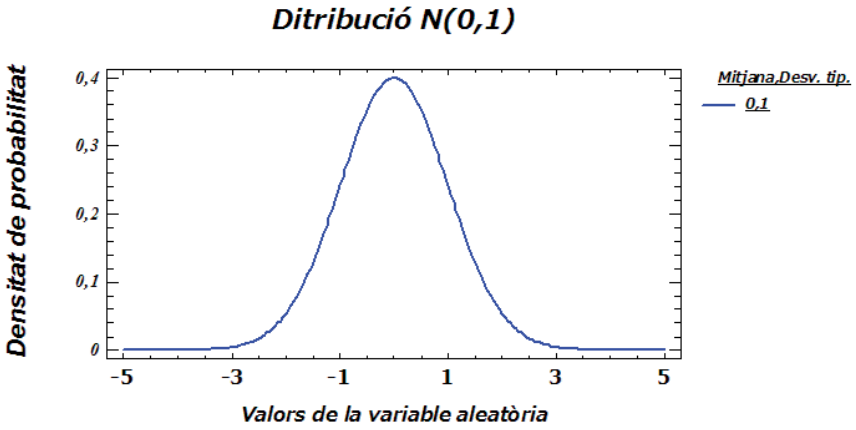


Figura 22

Com es pot comprovar, la funció de densitat té forma de campana (s'anomena *campana de Gauss*) i és simètrica respecte al valor de l'esperança (en aquest cas 0). També es pot comprovar que la densitat de probabilitat decreix a tots dos costats (les cues) i és molt més gran en els valors centrals propers a l'esperança 0. Aquest fet significa que els valors centrals tenen molta més probabilitat de donar-se que els valors que es troben a les cues.

Que passa si en una distribució normal augmenta o disminueix la desviació típica? Es comprovarà també mitjançant els gràfics de les respectives funcions de densitat de probabilitat.

Els gràfics següents mostren la distribució de probabilitat que segueixen les alçades de les persones en tres municipis diferents: en el primer municipi l'altura es distribueix segons una normal d'esperança 175 i desviació típica 1; la segona, segons una normal d'esperança 175 i desviació típica 3; i l'última, segons una normal d'esperança 175 i desviació típica 10.

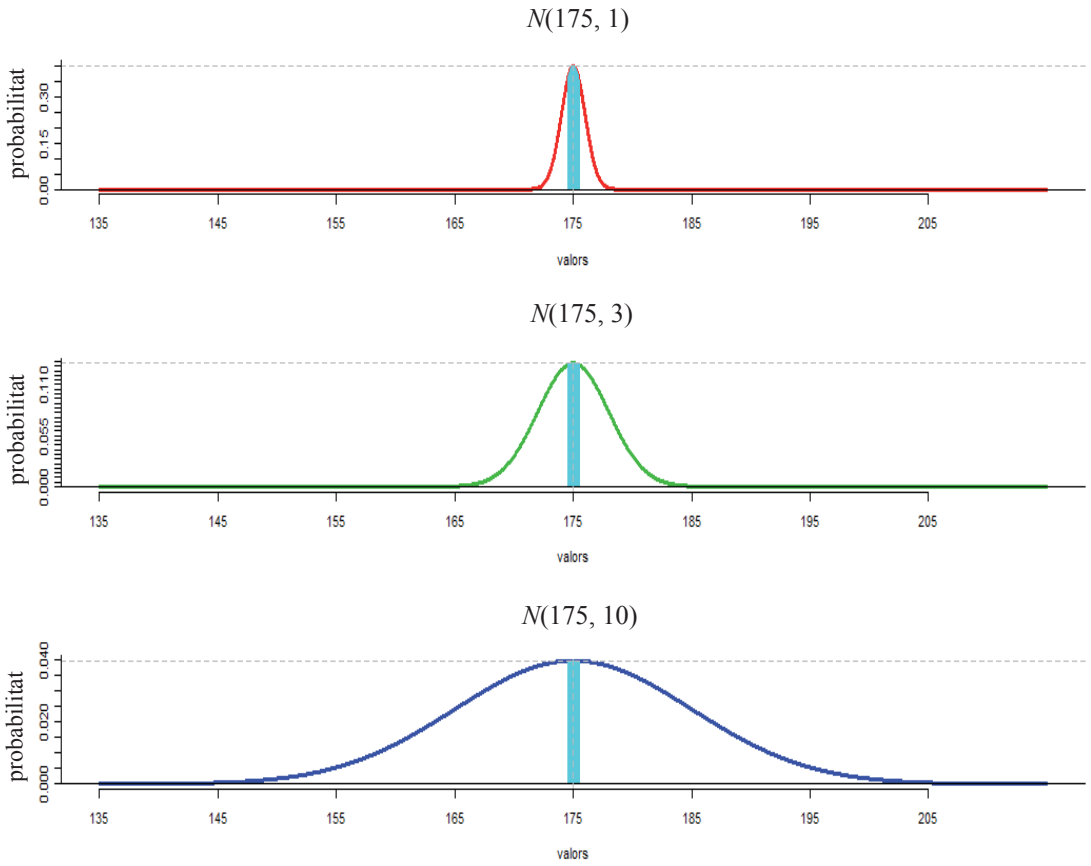


Figura 23

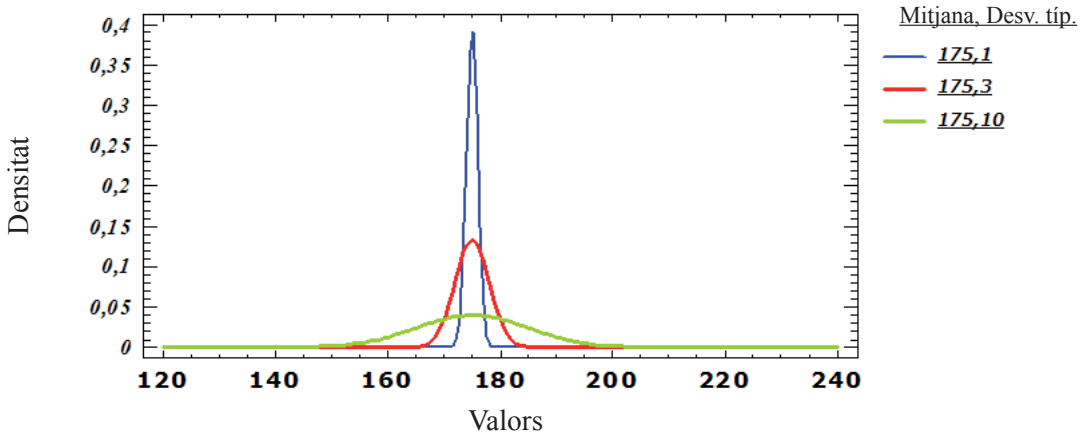


Figura 24

Com s'observa, és més probable trobar persones que mesuren valors propers a 175 en poblacions amb desviació típica 1 que en poblacions amb desviacions típiques 3 i 10.

Cal recordar que les probabilitats en les distribucions contínues es calculen mitjançant àrees. En el primer gràfic (figura 23) s'observa que la probabilitat de trobar persones amb una altura compresa entre 174,4 i 175,5 és més alta en la normal de variància 1 que en les altres dues. A més, aquesta probabilitat també és més alta en la normal amb variància 3 que en la de 10. Cal recordar que l'àrea compresa entre la funció de densitat i l'eix horitzontal és 1 en tots tres casos.

Així doncs, es pot concloure que a mesura que la desviació típica augmenta, la probabilitat de trobar persones que facen alçades pròximes al valor esperat (en aquest cas, 175 cm) disminueix. Una vegada resposta la pregunta «què ocorre si en una distribució normal varia la desviació típica», cal fer-se'n una altra: què ocorre si en una distribució normal es manté fixa la desviació típica però canvia el valor de l'esperança?

En els gràfics següents es representen tres distribucions normals que tenen la mateixa desviació típica però valor de l'esperança distint.

Concretament, les distribucions són $N(6, 1)$; $N(15, 1)$ i $N(25, 1)$, les quals corresponen als resultats obtinguts per l'alumnat de 14-16 anys en una prova de matemàtiques valorada de 0 a 30 punts en tres països diferents. És a dir, la variable aleatòria és X = nota obtinguda en una prova de matemàtiques.

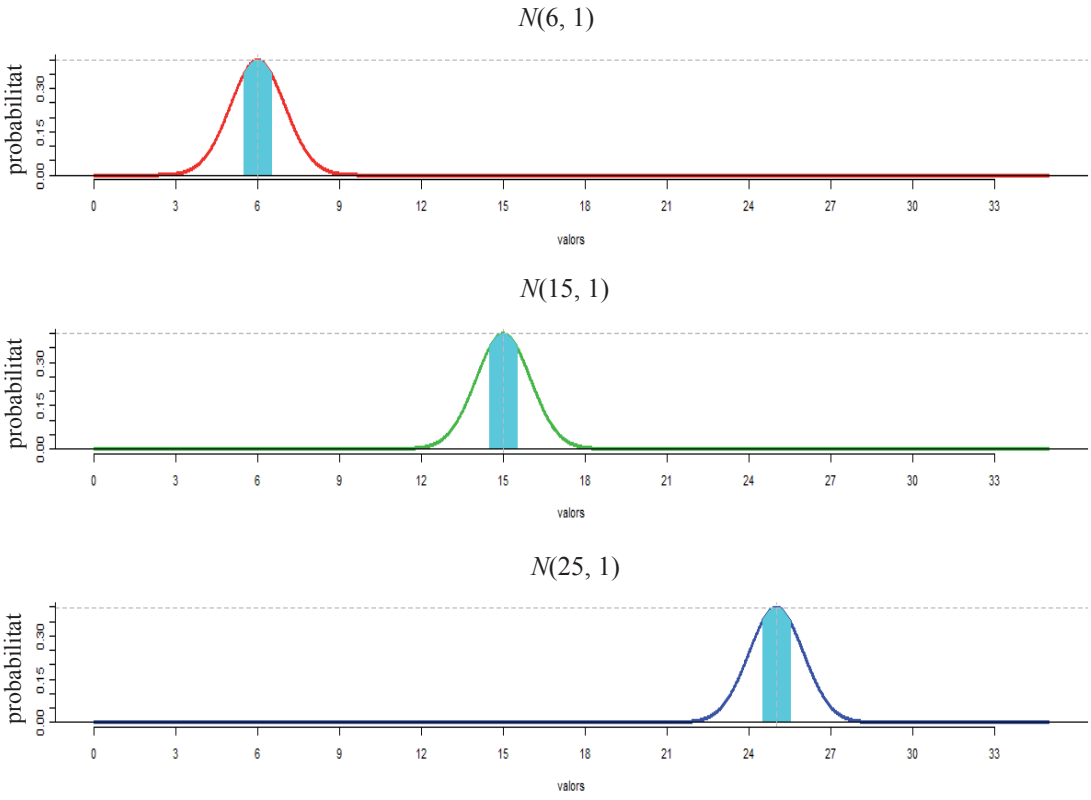


Figura 25

Llavors la nota obtinguda per l'alumnat del país 1 es distribueix $N(6, 1)$, la nota obtinguda per l'alumnat del país 2 es distribueix $N(15, 1)$ i la nota obtinguda per l'alumnat del país 3 es distribueix $N(25, 1)$.

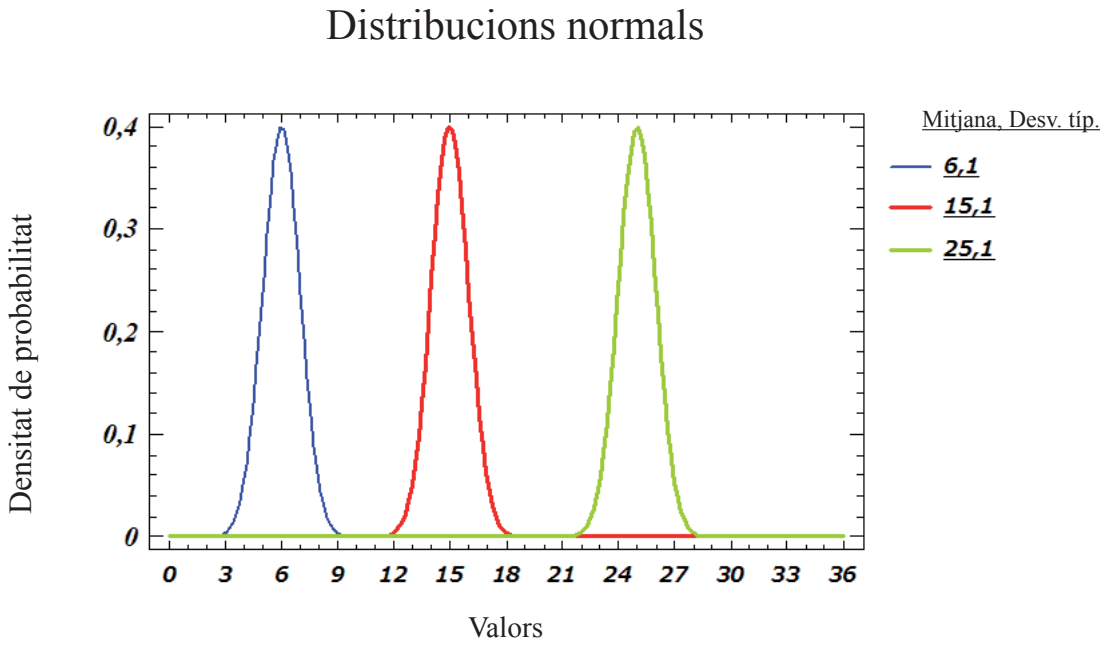


Figura 26

La representació gràfica de totes les distribucions és la mateixa (figura 26). Únicament varia el punt on se centren. Aquest punt és, evidentment, l'esperança de cadascuna de les distribucions: 6 per a la primera, 15 per a la segona i 25 per a la tercera.

Aquest fet significa que els valors que pren la variable aleatòria es distribueixen de la mateixa manera entorn de l'esperança. Així, per exemple, la probabilitat de trobar un alumne/a que haja fet la prova i haja obtingut una nota compresa entre 4 i 6 del país 1 és la mateixa que la de trobar un alumne/a que haja fet la prova i haja obtingut una nota compresa entre 14 i 16 del país 2, i és la mateixa que la de trobar un alumne/a que haja fet la prova i haja obtingut una nota compresa entre 24 i 26 del país 3.

És a dir:

$P(6-\alpha \leq X \leq 6+\alpha)$ en el país 1, és la mateixa que $P(15-\alpha \leq X \leq 15+\alpha)$ en el país 2 i la mateixa que $P(25-\alpha \leq X \leq 25+\alpha)$ en el país 3.

El teorema que tot seguit es presenta permet realitzar els càlculs de probabilitats de variables aleatòries que es distribueixen segons una distribució normal sense haver de realitzar els càlculs integrals. Així doncs, el teorema següent relaciona distribucions $N(\mu, \sigma)$ amb distribucions $N(0, 1)$.

Teorema de les transformacions lineals

Siguen $X \approx N(\mu, \sigma)$ i siguen $Y = a\,X + b$, on a i b són constants.
Aleshores $Y \approx N(a\mu + b, a \cdot \sigma)$.

Exemple 38

La qualificació obtinguda en un test per l'alumnat de primària d'una ciutat segueix una distribució normal de mitjana o esperança 5,5, i desviació típica 1. El comitè avaluador s'ha adonat que el qüestionari que feren els alumnes contenia diferents errades i decidex augmentar un 5% la nota de cada alumne/a. Com es distribuirà ara la qualificació del test?

La nova variable $Y = 1{,}05 \cdot X \rightarrow$ Com que $X \approx N(5{,}5,\ 1)$ aleshores, aplicant-hi el teorema, es té que $Y \approx N(5{,}775,\ 1{,}05)$.

Corol·lari

Si $X \approx N(\mu, \sigma)$, llavors la nova variable aleatòria $Z = \dfrac{X - \mu}{\sigma} \approx N(0, 1)$.

Exemple 39

En el cas de l'exemple anterior, si $X \approx N(5, 5, 1) \rightarrow Z = \dfrac{X - 5{,}5}{1} \approx N(0, 1)$.

7.7.2. Distribució normal tipificada

La distribució normal amb esperança o mitjana 0 i variància 1 es denomina *distribució normal tipificada*, $Z \approx N(0,1)$. La funció de distribució d'aquesta variable aleatòria es denota per Φ, on $\Phi(z) = P(Z \leq z)$, i s'utilitza per al càlcul de probabilitats amb taules, ja que les probabilitats acumulades per a Z estan tabulades. També s'hi pot utilitzar software informàtic. Al llarg del text s'empraran indistintament $\Phi(z)$ o $P(Z \leq z)$.

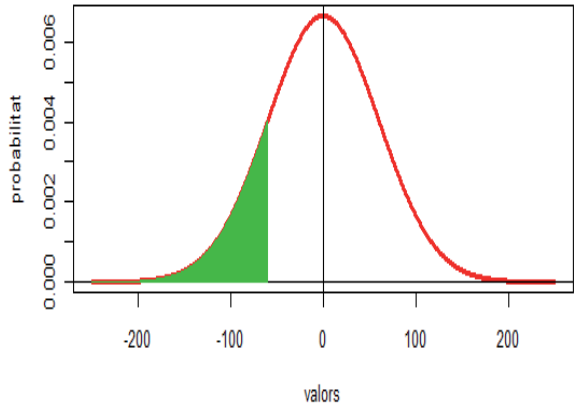
Important: per ser la distribució simètrica es compleix que (figura 27):

$$P(Z \leq z) = P(Z > -z) = 1 - P(Z \leq -z).$$

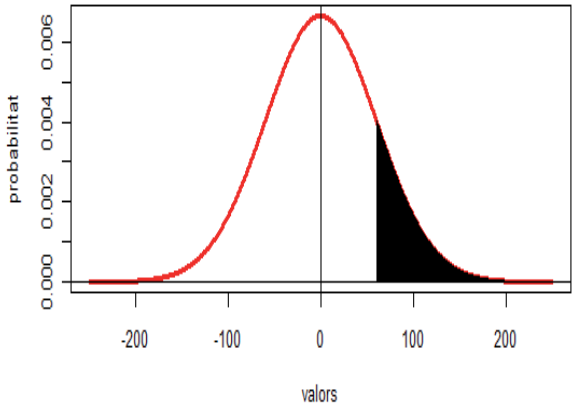
Així, $\Phi(-z) = 1 - \Phi(z)$.

És important familiaritzar-se amb el càlcul de probabilitats utilitzant la taula de la distribució $N(0, 1)$, ja que tots els càlculs de probabilitats es realitzen emprant aquestes taules.

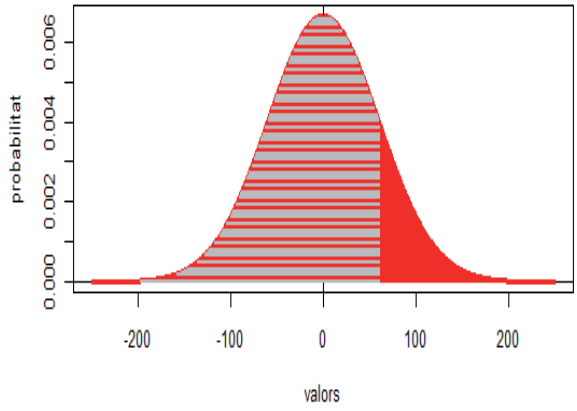
$F(-60) = P(X < -60)$



$P(X < -60) = P(X > 60)$



$P(X > 60) = 1 - P(X < 60)$



$P(X < -60) = \mathrm{P}(\mathrm{X} > 60) = 1 - P(X < 60)$

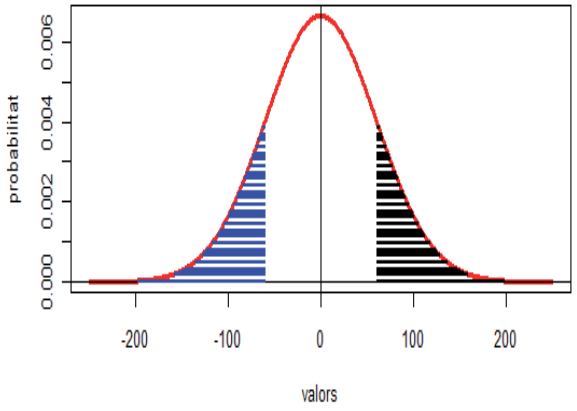


Figura 27

Exemple 40

Siga $Z \approx N(0, 1)$. Calcula: $P(Z \leq 2)$, $P(Z > 1.5)$, $P(1 \leq Z \leq 2{,}1)$, $P(Z \leq -0{,}65)$.

$P(Z \leq 2) = \Phi(2) = 0{,}9773$
$P(Z > 1{,}5) = 1 - P(Z \leq 1.5) = 1 - \Phi(1{,}5) = 1 - 0{,}9332 = 0{,}0668$
$P(1 \leq Z \leq 2{,}1) = P(Z \leq 2.1) - P(Z \leq 1) = \Phi(2.1) - \Phi(1) = 0{,}9821 - 0{,}8413 =$
$= 0{,}1408$
$P(Z \leq -0{,}65) = \Phi(-0{,}65) = 1 - \Phi(0{,}65) = 1 - 0{,}7422 = 0{,}2578$

Exemple 41

Siga $X \approx N(10, 5)$. Calcula: $P(X > 15)$, $P(X \leq 12)$, i $P(10 < X \leq 16)$.

Primerament cal tipificar la variable aleatòria X. És a dir, convertir-la en una normal de mitjana 0 i desviació típica 1.

$Z = \frac{Z - 10}{5} \rightarrow N(0, 1)$, i, per tant, la probabilitat anterior es pot escriure així:

$P(X > 15) = 1 - P(X \leq 15) = 1 - P(Z \leq \frac{15 - 10}{5}) = 1 - P(Z \leq 1) = 1 - 0{,}8413 = 0{,}1587.$

Anàlogament s'obté que:

$P(X \leq 12) = P(Z \leq \frac{12 - 10}{5}) = P(Z \leq 0{,}4) = 0{,}6554$

$P(10 < X \leq 16) = P(X \leq 16) - P(X \leq 10)$

$= P(Z \leq \frac{16 - 10}{5}) - P(Z \leq \frac{10 - 10}{5}) = P(Z \leq 1{,}2) - P(Z \leq 0) = 0{,}8849 - 0{,}5 = 0{,}3849.$

Gràficament, es pot comprovar en la figura següent:

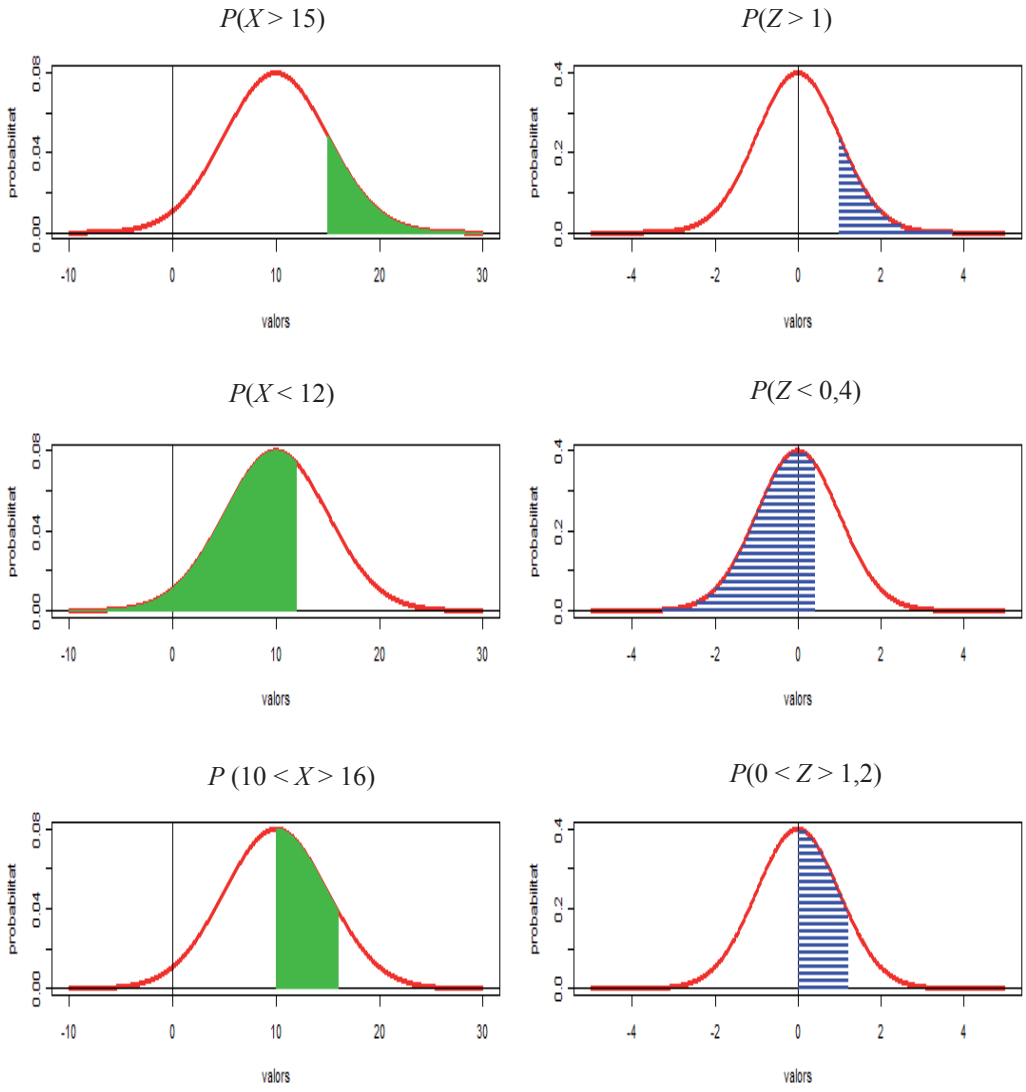


Figura 28

Exemple 42

L'estatura dels ciutadans d'una gran ciutat segueix una distribució normal de mitjana 1,7 i desviació típica 0,2.

- Se selecciona un ciutadà a l'atzar. Calcula la probabilitat que la seua altura siga superior a 1,95 m.
- Se selecciona a l'atzar un altre ciutadà entre els de talla superior a 1,65. Calcula la probabilitat que la seua estatura siga superior a 1,95.
- Se seleccionen a l'atzar 5 ciutadans, calcula la probabilitat que tres d'ells tinguen una estatura superior a 1,95.

1. Siga X = estatura. Aquesta variable aleatòria segueix una distribució $N(17, 0, 2)$.

   Es demana $P(X > 1{,}95) = 1 - P(X \leq 1{,}95)$. Com que la variable aleatòria no segueix una distribució $N(0, 1)$, és necessari tipificar-la per a poder utilitzar les taules.

   Així $Z = \dfrac{X - 1{,}7}{0{,}2}$ es distribueix $N(0, 1)$. Per tant, es demana $P(X > 1{,}95) = 1 - P(X \leq 1{,}95) = 1 - P(Z \leq \dfrac{1{,}95 - 1{,}7}{0{,}2}) = 1 - P(Z \leq 1{,}25) = 1 - 0{,}8944 = 0{,}1056$.

2. Aquest ciutadà està seleccionat entre els que mesuren més d'1,65. Per tant, s'hi demana una probabilitat condicionada:

$$P(X > 1{,}95/X > 1{,}65) = \frac{P(X > 1{,}95 \cap X > 1{,}65)}{P(X > 1{,}65)} = \frac{P(X > 1{,}95)}{P(X > 1{,}65)} =$$

$$\begin{cases} P(X > 1{,}95) = 0{,}1056; \\ P(X > 1{,}65) = 1 - P(X \leq 1{,}65) = 1 - P(Z \leq -0{,}25) = P(Z > 0{,}25) = \\ \qquad = 1 - (1 - P(Z \leq 0{,}25) = P(Z \leq 0{,}25) = 0{,}5987 \end{cases}$$

$$P(X > 1{,}95/X > 1{,}65) = \frac{0{,}1056}{0{,}5987} = 0{,}176$$

3. Se seleccionen 5 ciutadans a l'atzar. La probabilitat que un ciutadà tinga una estatura superior a 1,95 és de 0,1056.

   Si es defineix Y = nombre de ciutadans que mesuren més d'1,95 entre 5, és clar que Y es distribueix $Bi(5, 0, 1506)$.

   Es demana $P(Y = 3) = \binom{5}{3} 0{,}1056^3 \cdot 0{,}8944^2 = 0{,}0094$.

Exemple 43

Una de les fases en una oposició consisteix a realitzar una prova psicotècnica. Després de fer l'estudi sobre les puntuacions de 0 a 100, s'ha descobert que segueixen una distribució normal de mitjana 60 i desviació típica 15. Si per a passar a la fase següent s'eliminen el 10% de les persones que han obtingut més nota i el 10% de les que han obtingut menys nota, entre quines dues notes ha d'estar la qualificació obtinguda per un aspirant per a passar a la fase següent?

El que demana l'exemple és l' interval en què es troben el 80% de les notes centrals. D'una banda, cal calcular la nota a de manera que el 10% dels aspirants no l'hagen superada, és a dir, $P(X < a) = 0{,}1$. D'altra banda, cal trobar un nota b de manera que el 10% dels aspirants l'hagen superada, és a dir, $P(X > b) = 0{,}1$.

Per a trobar els valors de a cal tenir en compte que és necessari tipificar, ja que la distribució no és $N(0, 1)$.

Així:

$$P(X \leq a) = 0{,}1 \rightarrow P\left(\frac{a - 60}{15}\right) = 0{,}1.$$

Tenint en compte que en una distribució $N(0,1)$ la $P(Z \leq -1{,}28) = 0{,}1$, que es denota de la manera $Z_{0,1} = -1{,}28$, i igualant els valors $Z_{0,1} = \frac{a - 60}{15}$ s'obté:

$$\frac{a - 60}{15} = -1{,}28 \rightarrow a = 40{,}8.$$

D'una manera semblant per a trobar el valor de b:

$$P(X > b) = 0{,}1 \rightarrow P(X \leq b) = 1 - P(X > b) = 1 - 0{,}1 = 0{,}9$$

$$P(X \leq b) = 0{,}9 \rightarrow P\left(\frac{b - 60}{15}\right) = 0{,}9$$

Tenint en compte que en una distribució $N(0, 1)$ la $P(Z \leq 1{,}28) = 0{,}1$, que es denota de la manera $Z_{0,9} = 1{,}28$, i igualant els valors $Z_{0,9} = \frac{b - 60}{15}$ s'obté que:

$$\frac{b - 60}{15} = 1{,}28 \rightarrow b = 79{,}2.$$

En conseqüència, qualsevol aspirant amb una nota compresa entre [40,8, 79,4] passarà a la fase següent.

7.7.3. La distribució normal com una aproximació a la distribució binomial

Ja s'ha vist que el càlcul de probabilitats de la binomial (n, p) pot aproximar-se per la distribució de Poisson ($n \cdot p$) si n i p compleixen unes determinades condicions. D'una manera semblant, la distribució binomial pot aproximar-se per una distribució normal.

Així doncs, si una distribució X és Bi (n, p) i n és suficientment gran, es pot emprar l'aproximació normal en què $\mu = n \cdot p$ i $\sigma^2 = n \cdot p \cdot (1 - p)$. Aquesta aproximació és molt acurada quan $n \cdot p \geq 5$ $n \cdot (1 - p) \geq 5$.

Gràficament (figura 29) sembla bastant clar que com més gran és el valor de n més s'assembla la distribució binomial a una de normal.

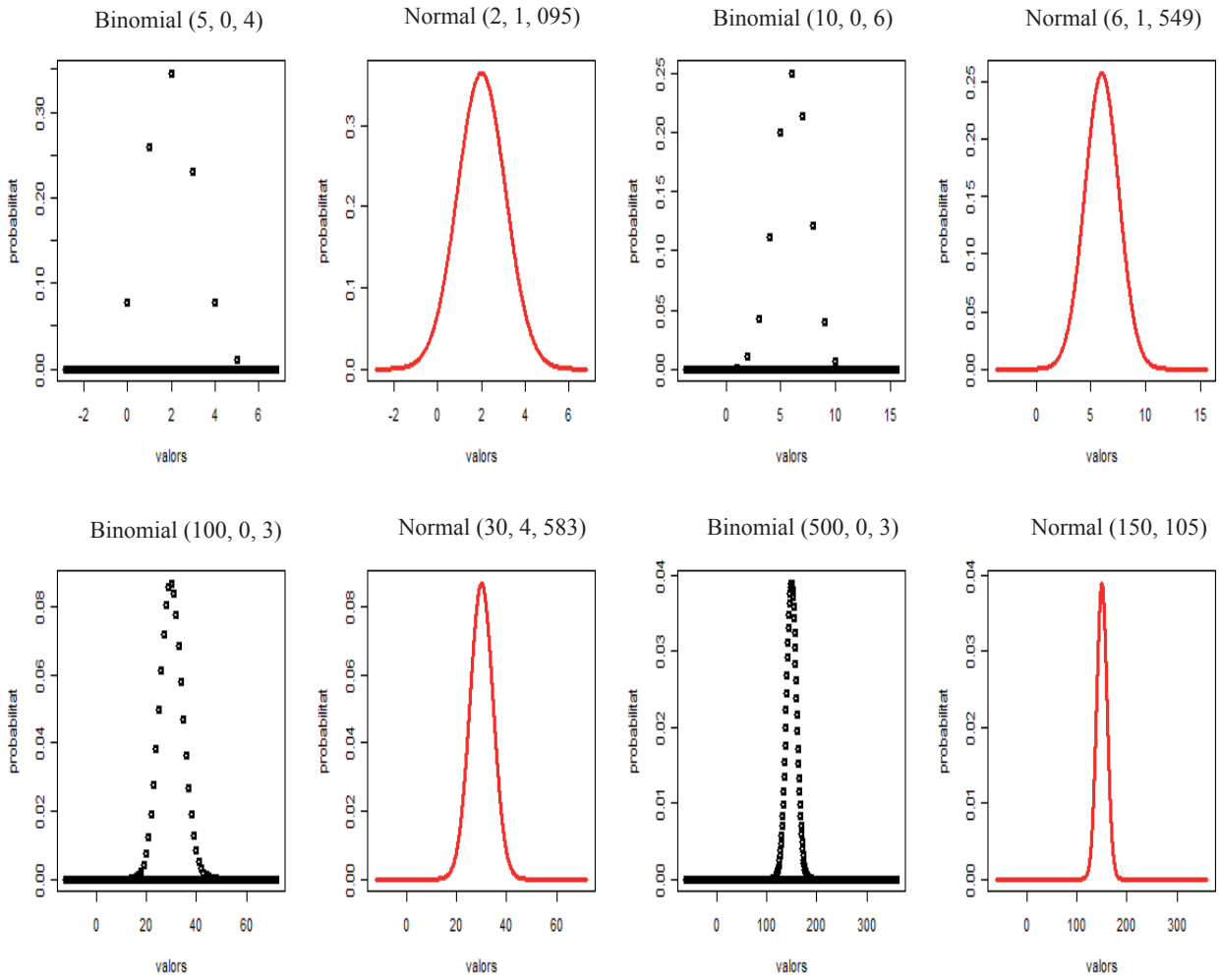


Figura 29

Com s'observa en els gràfics anteriors, en augmentar el nombre d'experiments (n) la distribució va assemblant-se cada vegada més a una distribució normal. Cal tenir en compte, però, que per ser la distribució binomial una distribució discreta, i la normal una distribució contínua, a l'hora de fer les aproximacions és necessari aplicar-hi una «correcció». Així, si es denota X = distribució binomial i X' = distribució normal que l'aproxima, llavors la correcció consisteix a assumir que $P(X = a) = P(a - 0{,}5 \leq X' \leq a + 0{,}5)$.

En el gràfic s'observa que en una distribució binomial (30, 0, 6) la $P(X = 17)$ és la mateixa que en la seua aproximació per la normal $N(18, 7{,}2)$, $P(17 - 0{,}5 \leq X' \leq 17 + 0{,}5)$.

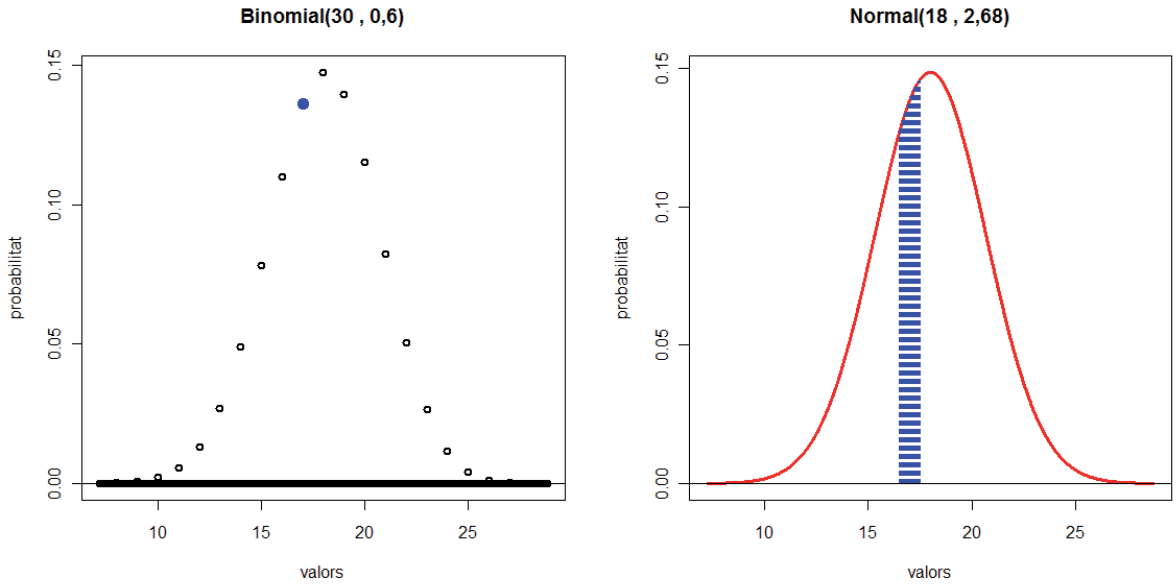


Figura 30

Numèricament, es pot calcular emprant software estadístic:

Valors(x_i)	Binomial $P(X = x_i)$	Normal $P(x_i - 0{,}5 < X < x_i + 0{,}5)$	Diferència
15	7.831221e – 02	7.969387e – 02	1.381658e – 03
16	1.101265e – 01	1.123280e – 01	2.201428e – 03
17	1.360387e – 01	1.380144e – 01	1.975729e – 03
18	1.473752e – 01	1.478211e – 01	4.458439e – 04
19	1.396186e – 01	1.380144e – 01	–1.604236e – 03
20	1.151854e – 01	1.123280e – 01	–2.857403e – 03
21	8.227527e – 02	7.969387e – 02	–2.581401e – 03
22	5.048710e – 02	4.928696e – 02	–1.200133e – 03
23	2.634109e – 02	2.657081e – 02	2.297196e – 04

Valors(x_i)	Binomial $P(X = x_i)$	Normal $P(x_i - 0{,}5 < X < x_i + 0{,}5)$	Diferència
24	1.152423e – 02	1.248641e – 02	9.621778e – 04
25	4.148722e – 03	5.114733e – 03	9.660103e – 04
26	1.196747e – 03	1.826220e – 03	6.294734e – 04
27	2.659437e – 04	5.683541e – 04	3.024103e – 04
28	4.274096e – 05	1.541737e – 04	1.114327e – 04
29	4.421478e – 06	3.645155e – 05	3.203007e – 05
30	2.210739e – 07	7.511459e – 06	7.290385e – 06
31	0.000000e + 00	1.349033e – 06	1.349033e – 06
32	0.000000e + 00	2.111533e – 07	2.111533e – 07
33	0.000000e + 00	2.880289e – 08	2.880289e – 08
34	0.000000e + 00	3.423917e – 09	3.423917e – 09
35	0.000000e + 00	3.546865e – 10	3.546865e – 10
36	0.000000e + 00	3.201728e – 11	3.201728e – 11
37	0.000000e + 00	2.518430e – 12	2.518430e – 12
38	0.000000e + 00	1.726397e – 13	1.726397e – 13
39	0.000000e + 00	1.032507e – 14	1.032507e – 14
40	0.000000e + 00	5.551115e – 16	5.551115e – 16
41	0.000000e + 00	0.000000e + 00	0.000000e + 00
42	0.000000e + 00	0.000000e + 00	0.000000e + 00
43	0.000000e + 00	0.000000e + 00	0.000000e + 00
44	0.000000e + 00	0.000000e + 00	0.000000e + 00
45	0.000000e + 00	0.000000e + 00	0.000000e + 00

Taula 3

Com s'observa en la taula 3, els valors de la binomial i de la normal amb correcció s'ajusten molt. El gràfic següent mostra que els valors de les diferències poden considerar-se quasi menyspreables per estar molt pròxims al zero.

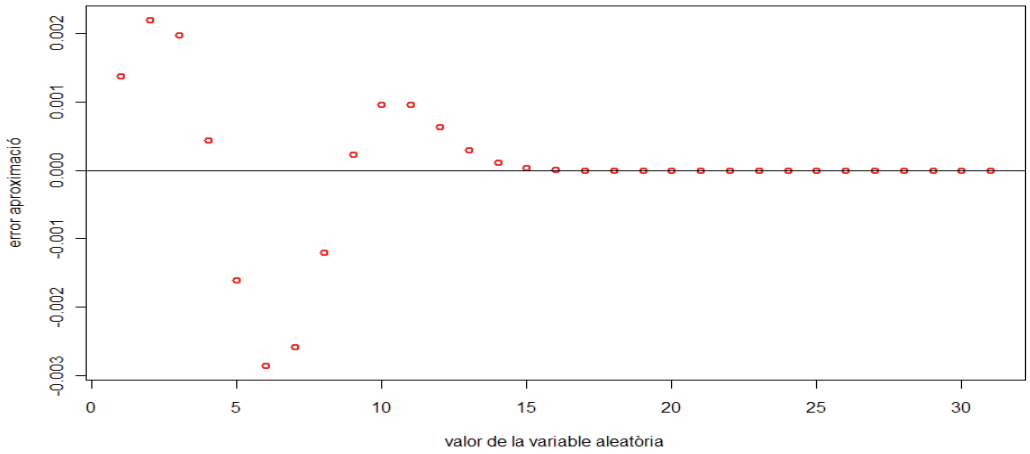


Figura 31

Exemple 44

Un treballador d'una agència d'assegurances telefona a diferents persones per oferir el seu producte. L'empresa sap que per a aquest treballador el 15% de les trucades tenen com a resultat un contracte d'una assegurança. Si per al mes que comença el treballador té programades 200 telefonades, quina és la probabilitat que aconseguisca almenys 40 contractes? I que n'aconseguisca 25?

Si s'assigna X = nombre de contractes aconseguits, sembla clar que $X \approx Bi(200, 0,15)$ i l'exercici demana $P(X > 40)$ i $P(X = 25)$. Com que el valor de n és molt gran, es pot emprar l'aproximació per la distribució normal $X' \approx N\ (200 \cdot 0{,}15, \sqrt{200 \cdot 0{,}15 \cdot 0{,}85})$, és a dir, $X' \approx N(30, 5{,}05)$. Aleshores, $P(X = 25) = P(24{,}5 \leq X' \leq 25{,}5) =$

$$= P\left(Z \leq \frac{25{,}5 - 30}{5{,}05}\right) - P\left(Z \leq \frac{24{,}5 - 30}{5{,}05}\right)$$

$$= \Phi(-0{,}89) - \Phi(-1{,}09)$$

$$= 1 - \Phi(0{,}89) - 1 \ - \Phi(1{,}09)$$

$$= \Phi(1{,}09) - \Phi(0{,}89) = 0{,}04838781.$$

Cal notar que el valor donat per la distribució binomial és 0,05080383 i, per tant, l'aproximació és molt acurada.

Pel que fa a la primera pregunta, l'exercici demana:

$$P(X > 40) = P(X' > 40) = 1 - P(X' \leq 40{,}5) =$$

$$= 1 - P\left(Z \leq \frac{40{,}5 - 30}{5{,}05}\right)$$
$$= 1 - \Phi(2{,}07) =$$
$$= 0{,}0192.$$

En aquest cas el valor que dóna la distribució binomial és 0,02199927.

Nota

La distribució normal també pot emprar-se per a aproximar la distribució de Poisson quan el paràmetre és molt gran. La mitjana i la desviació típica de la distribució normal són «heretades» de la de Poisson. El gràfic següent mostra les funcions de distribució de les tres variables aleatòries en un dels casos en què podria fer-se l'aproximació.

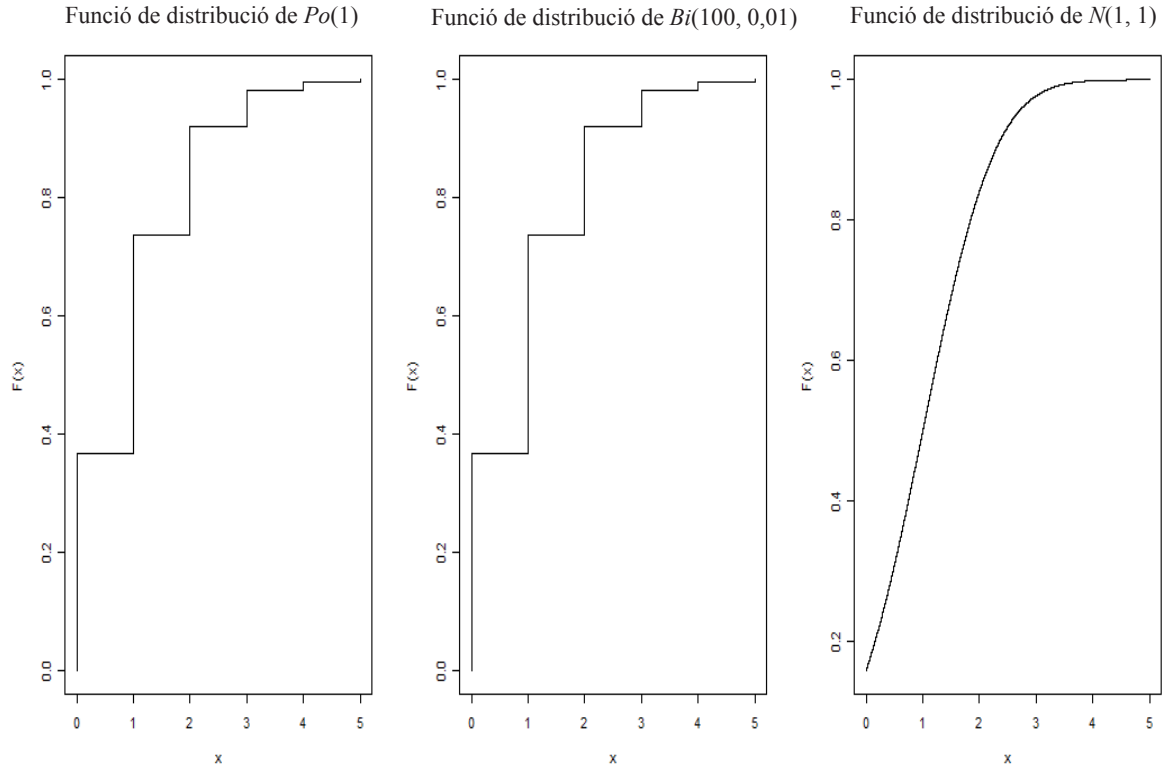


Figura 32

7.7.4. Teorema del límit central

Les distribucions binomial i de Poisson es poden aproximar per la distribució normal, com ja s'ha comentat al llarg del text. Analitzant amb més detall totes dues distribucions, s'observa que la primera es pot considerar com una suma de variables aleatòries independents de Bernoulli amb probabilitat d'èxit constant, i la segona, com una suma de distribucions de Poisson independents (tan sols cal dividir l'interval de temps o espai en intervals més petits i aplicar-hi la propietat de la suma de distribucions de Poisson). Per tant, d'alguna manera, la distribució normal s'ha emprat per a aproximar distribucions suma de variables aleatòries independents. Però aquest privilegi tan sols és gaudit per les distribucions binomial i de Poisson? La resposta és que no, i la raó, el teorema del límit central.

Així doncs, el teorema del límit central estudia el comportament de la suma (o la mitjana) de variables aleatòries. És a dir, el teorema del límit central afirma que la distribució d'una suma molt gran de variables aleatòries s'aproxima a una distribució normal. D'aquesta manera, per una banda, aporta a l'estadística un resultat crucial per a l'estudi de la distribució asimptòtica de moltes variables aleatòries. Com es comprovarà en els capítols posteriors, serà de gran utilitat per als contrastos d'hipòtesi i els intervals de confiança. Per altra banda, el teorema del límit central proporciona una explicació teòrica fonamentada en un fenomen habitual en els experiments reals: les variables presenten en moltes ocasions una distribució empírica aproximadament normal.

Encara que com qualsevol teorema important de les matemàtiques s'ha anat construint al llarg de la història –i existeix més d'una versió d'aquest teorema depenent de les exigències de les hipòtesis–, un dels enunciats més comuns és el que es presenta tot seguit.

Teorema del límit central

Donades X_1, X_2, ..., X_n variables aleatòries independents i idènticament distribuïdes amb esperança μ i variància σ^2, i n és suficientment gran, la variable aleatòria $X_1 + X_2 + ... + X_n$ té una distribució normal d'esperança $n \cdot \mu$ i variància $n \cdot \sigma^2$.

És a dir:

$$X_1 + X_2 + ... + X_n \approx N(n \cdot \mu , \sqrt{n} \cdot \sigma)$$

Cal remarcar que aquest teorema també admet la seua versió equivalent:

$$\frac{X_1 + X_2 + ... + X_n}{n} = \overline{X} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Exemple 45

El volum de vendes anuals dels comercials d'un país en milions d'euros segueix una distribució normal de mitjana 1,25 i desviació típica 0,75. S'escullen a l'atzar 50 comercials del país. Quina és la probabilitat que el volum de vendes de tots plegats siga almenys de 65 milions d'euros? Quina és la probabilitat que la mitjana de vendes siga inferior a 1,5 milions d'euros?

Es defineix X_1 = volum de vendes de la primera comercial seleccionada, X_2 = volum de vendes del segon comercial seleccionat, ..., X_{50} = volum de vendes del 5è comercial seleccionat.

La primera pregunta demana $P(X_1 + X_2 + ... + X_{50} \geq 65)$. Aleshores cal saber, en primer lloc, com es distribueix $X_1 + X_2 + ... + X_{50}$. Sembla clar que es compleixen les condicions per a poder aplicar el teorema del límit central, ja que cadascuna de les 50 variables aleatòries es distribueixen de la mateixa manera i són independents. En conseqüència:

$$X_1 + X_2 + ... + X_{50} \approx N(50 \cdot 1, 25, \sqrt{50} \cdot 0{,}75) = N(62, 5, 5, 30)$$

$$P(X_1 + X_2 + ... + X_{50} \geq 65) = 1 - P(X_1 + X_2 + ... + X_{50} < 65) =$$

$$= 1 - P\left(Z \leq \frac{65 - 62{,}5}{5{,}30}\right) = 1 - (0{,}47) = 0{,}3197.$$

Pel que fa a la segona pregunta, l'exercici demana $P\left(\frac{X_1 + X_2 + ... + X_{50}}{50} < 1{,}5\right)$ és a dir, demana $P(\overline{X} < 1{,}5)$.

Pel teorema del límit central se sap que $\overline{X} \approx N(1{,}5, \ 0{,}11)$. Aleshores,

$$P(\overline{X} < 1{,}5) = P\left(Z \leq \frac{1{,}5 - 1{,}25}{0{,}11}\right) = (2{,}27) = 0{,}9884.$$

7.8. Problemes proposats

En aquest epígraf es plantejaran un conjunt de problemes per a la resolució dels quals és necessari conèixer la teoria desenvolupada al llarg de la unitat.

Exercici 1

Sabem que un comercial de telefonia aconsegueix fer una venda el 12% de les vegades que ho intenta. També sabem que en un dia fa 20 telefonades per hora. Hem decidit contractar aquest sistema per a la promoció del producte que volem vendre. Calcula:

a) La mitjana de vendes per hora.
b) La variància.
c) La probabilitat que el nombre de vendes en una hora estiga entre 1 i 4, ambdós inclosos.

Exercici 2

Per a diagnosticar una malaltia laboral, es disposa d'una prova que dóna positiu en el 30% dels malalts, però no dóna mai positiu si el individu està sa. Per assegurar-nos-en, aquest diagnòstic s'aplica a tota la població, fins a tres vegades si abans no ha donat positiu. Quina proporció de malalts queda sense detectar?

Exercici 3

En una fase d'una oposició, els aspirants han de desenvolupar un tema. Els opositors han d'escollir-ne un dels cinc que el tribunal tria de manera aleatòria d'un conjunt de 72 temes. Si una persona es presenta a l'oposició i se n'ha estudiat 30, calcula:

a) Quina és la probabilitat que s'haja estudiat almenys un dels cinc temes que ha extret el tribunal? I que se n'haja estudiat almenys dos dels cinc?
b) Quina és la probabilitat que no se n'haja estudiat cap dels cinc?

Exercici 4

Per fer un estudi de qualitat dels serveis, una empresa arreplega diàriament 3.600 enquestes que omplin els seus clients, de les quals una proporció no estan completes i cal retirar-les. La distribució d'enquestes defectuoses segueix una distribució normal, la mitjana de la qual és de 144 i la desviació típica, de 11,75. Determina:

a) Probabilitat que ens trobem més de 150 enquestes defectuoses un dia determinat.
b) Probabilitat que el nombre d'enquestes defectuoses en un dia qualsevol estiga comprès entre 140 i 160.

Exercici 5

El nombre de transferències bancàries que una oficina realitza per setmana es distribueix com una variable normal de mitjana 1.000 i desviació típica, 180.

a) Calcula la probabilitat que en una setmana determinada:
El nombre de transferències realitzades siga superior a 850.
El nombre de transferències realitzades estiga entre 1.000 i 1.200.
b) Si triem cinc setmanes aleatòriament, calcula la probabilitat que en més de dues setmanes el nombre de transferències realitzades siga inferior a 850.
c) Si sabem que la comissió que cobren per cada transferència és de 2,50 €, explica com es distribueix la variable que mesura els ingressos setmanals de l'oficina deguts a les transferències, i calcula els paràmetres d'aquesta variable.

Exercici 6

El departament d'una universitat decideix comprar una estació de treball. Se sap que la tecnologia fa que es produïsquen innovacions de manera aleatòria de tal manera que, per terme mitjà, ix un nou model que deixa obsolets els ja existents cada set mesos.

Quina és la probabilitat que la nova màquina es quede obsoleta al llarg d'un període de temps comprès entre sis mesos i un any?

Exercici 7

La probabilitat que en una gran empresa un missatge es reba amb errors és de 0,001. Si en un mes es reben 2.000 missatges, calcula la probabiliat de rebre'n dos amb errors.

TEMA 8

Introducció a la inferència estadística. Estimació puntual

Objectius tema 8

- Conèixer el procés de la inferència estadística.
- Conèixer i saber diferenciar els conceptes *població* i *mostra*.
- Conèixer els principals tipus de mostreig.
- Diferenciar entre *paràmetre* i *estadístic*.
- Comprendre el concepte *distribució* en el mostreig d'un estadístic.
- Conèixer el concepte *estimació puntual*.

8.1. Introducció

Tot estudi estadístic té per objectiu analitzar una determinada característica en una població. Seria ideal poder observar i mesurar la dita característica en tots els individus encara que sovint aquest plantejament global no és possible.

Imaginem, per exemple, els estudis estadístics següents:

- Es vol conèixer l'alçada mitjana de la població major d'edat de la nostra ciutat. Requereix un temps i unes despeses econòmiques excessives prendre la mesura a tots els ciutadans.
- Per a fer un control de qualitat del procés de fabricació en una factoria de mistos, caldria conèixer-ne la proporció de productes defectuosos. Seria un desgavell provar cadascun dels elements, ja que es destruiria la producció.

En ambdós casos, cal estudiar únicament una part de la població que ha de ser seleccionada amb criteris adequats per a assegurar-se que es tracta d'una mostra representativa de la població total i de les seues particularitats internes. Aquests criteris es presentaran en l'apartat següent.

En la pràctica és molt freqüent haver de recórrer a una mostra per inferir dades d'una població per alguns dels motius següents:

- La població és excessivament nombrosa: tota la població major d'edat de la nostra ciutat.
- La població és molt difícil o impossible de controlar: el nombre d'individus de certa espècie d'aus en un determinat parc natural.
- El procés de mesurament és destructiu: la comprovació de mistos defectuosos en una producció.
- Es volen conèixer els resultats ràpidament i es tardaria massa a prendre les mesures de tots: el sondeig d'opinió electoral.

8.2. Població i mostra. Tipus de mostreig

Població o univers és el conjunt de tots els individus objecte del nostre estudi. Mostra és un subconjunt extret de la població. El seu estudi serveix per a inferir característiques de tota la població.

L'elecció de la mostra es diu *mostreig*. Un mètode molt eficaç per a aconseguir una mostra representativa és que aquesta haja estat triada aleatòriament, és a dir, a l'atzar. Per contra, si l'elecció és subjectiva, els prejudicis de qui fa l'elecció es

projecten en el resultat de la mostra, la qual reflectirà el que aquesta persona creu que és la realitat i no s'aconseguiran mostres representatives. Recordem el que s'havia esmentat en el primer capítol: en les eleccions americanes de 1936, en els quals va guanyar Roosevelt, una revista va fer una enquesta d'intenció de vot a més de quatre milions dels seus lectors i es va equivocar en el pronòstic. Una altra enquesta realitzada només a 4.500 persones va anunciar l'èxit de Roosevelt amb molta exactitud.

La raó és que en el primer cas la mostra no era representativa de la societat americana, perquè tots eren lectors d'una mateixa revista, mentre que en les 4.500 persones de la segona mostra estaven ben representats tots els estaments i les ideologies d'aquesta societat.

Es diu que un mostreig és aleatori quan tots els individus de la mostra s'elegeixen a l'atzar, de manera que tots els individus de la població tenen, a priori, la mateixa probabilitat de ser triats.

Al llarg d'aquest epígraf se suposarà que es té una població de N elements de la qual es vol extraure una mostra de n elements. Es comentarà a continuació com es pot realitzar el mostreig perquè proporcione mostres representatives.

Es prendrà com a exemple l'extracció d'una mostra de 5 alumnes d'una classe de 36.

8.2.1. Mostreig aleatori simple

És el tipus de mostreig més senzill i en el qual es basen tots els altres.

Descripció	Exemple
- Numerem els elements de la població des de l'1 fins a N. - Fem un sorteig per a seleccionar els n elements que ha de contenir la mostra.	- Numerem els alumnes seguint l'ordre alfabètic dels cognoms. - Amb la calculadora, anotem una seqüència de nombres aleatoris de dues xifres i agafem els primers cinc nombres inferiors o iguals a 36. ~~91~~, 31, 21, ~~56~~, ~~48~~, 10, ~~89~~, 34, 11 La mostra està formada per l'alumnat que es correspon amb els nombres 31, 21, 10, 34 i 11.

Aquest sorteig pot fer-se de moltes maneres: extraure paperetes d'un caixó, extraure boles numerades d'una urna, etc. També podem ajudar-nos de la generació de nombres aleatoris fent servir la calculadora (tecla RAN#) o l'ordinador.

Si després de cada extracció l'individu seleccionat de la població pot tornar a ser-ho, es diu que és un mostreig aleatori amb reemplaçament. En cas contrari, és un mostreig aleatori sense reemplaçament.

8.2.2. Mostreig aleatori sistemàtic

La tria dels elements assegura una manera uniforme d'escollir els elements al llarg de la llista de la població.

Descripció	Exemple
- Numerem els elements de la població des de l'1 fins a N. - Triem per sorteig un individu qualsevol $N_{0.}$ - Calculem k, l'enter més proper al coeficient d'elevació $\frac{N}{n}$. - Seleccionem l'individu N_0 i els següents de k en k a partir d'ell. Cal tenir en compte que en sobrepassar N hem de continuar pel començament de la llista.	- Numerem els alumnes seguint l'ordre alfabètic dels cognoms. - Obtenim per sorteig un número. Imaginem que és el 21. - Trobem el coeficient d'elevació: $\frac{N}{n} = \frac{36}{5} = 7{,}2\ k = 7.$ - Comencem pel 21 i anem triant la mostra, agafant els elements de 7 en 7, presentem amb detall l'estratègia en arribar a 36. $[21] \underset{+7}{\rightarrow} [28] \underset{+7}{\rightarrow} [35] 36, 1, 2, 3, 4, 5 [6] \underset{+7}{\rightarrow} [13]$ La mostra està formada per l'alumnat que corresponen als nombres 21, 28, 35, 6 i 13.

Aquest procediment es pot realitzar sempre que en la numeració dels individus de la població no sospitem cap regularitat. No seria adient per a fer un estudi relacionat amb el trànsit agafar els dies amb periodicitat setmanal, mensual, etc.

8.2.3. Mostreig aleatori estratificat

El mostreig aleatori estratificat proporcional pressuposa que la població està formada per grups diferenciats que es denominen *estrats*. En el cas de l'afixació proporcional aquests grups estan representats en la mostra en la mateixa proporció numèrica que ho estan en la població.

Així, es considera que la població de N individus està formada per diferents estrats de grandària N_1, N_2, ... N_t que compleixen la condició: $N_1 + N_2 + ... + N_t = N$. Cal trobar el nombre d'individus n_1, n_2, ..., n_t que han de ser seleccionats per mostreig aleatori de cadascun dels estrats, de manera que $n_1 + n_2 + ... + n_t = n$. Per a calcular-los plantegem la proporció següent:

$$\frac{N}{n_1} = \frac{N_2}{n_2} = ... = \frac{N_t}{n_t} = \frac{N}{n}.$$

Descripció	Exemple
- Identifiquem els estrats en què dividirem la població i calculem N_1, N_2, ... N_t. - Calculem la grandària de les mostres en els estrats n_1, n_2, ..., n_t, mitjançant la proporció: $\frac{N}{n_1} = \frac{N_2}{n_2} = \ldots = \frac{N_t}{n_t} = \frac{N}{n}$. - En cada mostra de cada estrat es trien els individus per mostreig aleatori.	- Imaginem que els alumnes estan diferenciats en 21 xiques i 15 xics: $N_1 = 21,\ N_2 = 15.$ - Per a calcular n_1, n_2 resoldrem les proporcions: $\frac{N_1}{n_i} = \frac{N}{n}$, i obtenim $n_1 = 3$, $n_2 = 2$. La mostra està formada per 3 xiques i 2 xics, que triarem dins de cada grup mitjançant qualsevol dels procediments anteriors.

Hi ha altres mètodes d'afixació no proporcional en el mostreig estratificat però no seran abordats per tenir aquest text un caràcter introductori.

8.2.4. Mostreig aleatori per conglomerats

El mostreig aleatori per conglomerats pressuposa que la població està formada per grups diferenciats que s'anomenen *conglomerats*. Es considera que qualsevol és una mostra representativa de la població, ja que en cadascun està proporcionalment reflectida la diversitat de les seues característiques.

Podem utilitzar com a mostra qualsevol conglomerat.

Per exemple, si estem fent un treball d'estudi de mercat sobre l'acceptació de cert producte genèric en una urbanització d'habitatges que és molt homogènia (tots els habitatges tenen el mateix preu, es van construir simultàniament, els seus habitants tenen costums professionals i poder adquisitiu semblants, etc.), podem considerar que cada carrer d'aquesta urbanització és un conglomerat. Així, es pot escollir aleatòriament un carrer per a fer l'enquesta.

Per a concloure, es pot dir que els estrats són grups homogenis internament i que entre tots reflecteixen l'heterogeneïtat de la població, mentre que els conglomerats són homogenis entre si, però dins de cadascun està representada l'heterogeneïtat de la població.

8.3. Inferència. Paràmetres i estadístics

Una vegada seleccionats els elements de la mostra, es fa un treball d'estadística descriptiva, i es calculen els paràmetres de la mostra que es necessiten (mitjana aritmètica, variància, etc.). Basant-se en aquests resultats, es poden obtenir informació i conclusions de la característica de la població que es vol conèixer. Aquest darrer treball és l'objectiu de la inferència estadística.

Així doncs, la inferència estadística pot definir-se com la part de l'estadística que té per objecte el desenvolupament de tècniques que permeten conèixer o comprovar el valor dels paràmetres d'una població, a partir de les dades obtingudes d'una petita part que se n'extrau, que hem denominat *mostra*.

La fiabilitat d'aquestes deduccions es mesura en termes probabilístics, és a dir, tota afirmació en inferència va acompanyada de la seua probabilitat d'encert. Per la qual cosa es pot veure que s'arriba a les conclusions de la inferència aplicant la teoria de la probabilitat, de la qual ja s'han vist alguns dels resultats en els temes anteriors.

La inferència estadística té dues grans branques:

- Estadística inductiva o estimació, l'objecte de la qual és estimar el valor dels paràmetres de la població. Pot ser:
  - Estimació puntual, que desenvoluparem en els darrers apartats d'aquest tema.
  - Estimació per intervals, que desenvoluparem en el tema 9.
- Contrast d'hipòtesi, l'objecte de la qual és comprovar mitjançant mètodes matemàtics, hipòtesis realitzades sobre el valor d'algun paràmetre de la població. Desenvoluparem aquesta branca en el tema 10.

Si reprenem els exemples que plantejàvem en començar el tema:

- Volem intuir el valor de la mitjana de l'alçada de la població de la nostra ciutat, és a dir, la mitjana poblacional, que denotarem per μ. Per a això, triarem un grup de persones que reflectisquen la diversitat de la població. Serà, per tant, una mostra representativa. Calcularem la mitjana de l'alçada d'aquests individus, és a dir, calcularem la mitjana mostral, que denotarem per $\bar{x}$.
- En una factoria, podrem esbrinar la proporció de mistos defectuosos en la producció total, és a dir, la proporció poblacional, que denotarem per p, seleccionant un nombre prefixat d'objectes que extraurem de la cadena de producció i que conformen la mostra. D'aquests mistos, comptarem la proporció que n'hi ha de defectuosos, que serà la proporció mostral i denotarem per **p.**

No obstant això, cal remarcar que $\bar{x}$ no té exactament el valor de μ, ni **p** coincideix exactament amb el valor de p.

Així, quan treballem amb mostres, caldrà diferenciar aquests paràmetres calculats amb els individus seleccionats –els quals denominem *paràmetres estadístics* o simplement *estadístics*–, dels paràmetres reals corresponents a la població –els quals volem esbrinar, mitjançant les tècniques de la inferència que estudiarem–, i que denotem com a *paràmetres poblacionals* o simplement *paràmetres*.

Així, denotarem, en general, per θ, el paràmetre d'una població que cal estimar i que serà, segons el cas, la mitjana μ, la variància σ^2, i en el cas de poblacions de Bernoulli, la p, entre d'altres. Hem comentat que per a trobar aquest valor triarem una mostra representativa de grandària n i amb aquestes dades, en calcularem la mitjana aritmètica $\bar{x}$, la variància S^2_x, o la proporció **p** en el cas d'una població de Bernoulli, segons siga el valor que cal estimar.

En general, les mesures estudiades en l' estadística descriptiva dels primers temes són exemples d'estadístics (si els considerem com a funcions), ja que els seus valors depenen de les dades obtingudes en les mostres. Aquest paràmetre, que s'obté a partir del càlcul amb els valors coneguts d'una mostra, és el que direm *estadístic* i el denotarem en aquest tema, de manera genèrica, per la lletra T.

Intentarem explicar, ara, que aquests estadístics són variables aleatòries, ja que els valors mostrals també ho son. Per exemple, imaginem totes les mostres de cinc elements que podem obtenir de la població que ens havíem plantejat en l'exemple 1, on volíem conèixer l'alçada de la població dels majors d'edat de la nostra ciutat.

Per a això, seleccionarem cinc persones a les quals mesurarem l'alçada en centímetres. Si considerem les mesures d'unes mostres qualssevol, podrien ser, per exemple:

$(x_1, x_2, x_3, x_4, x_5) = (162, 173, 191, 182, 171) \rightarrow \bar{x} = 175{,}8$
$(x_1, x_2, x_3, x_4, x_5) = (155, 172, 190, 180, 175) \rightarrow \bar{x} = 174{,}4$
$(x_1, x_2, x_3, x_4, x_5) = (167, 182, 152, 152, 175) \rightarrow \bar{x} = 165{,}6$
$(x_1, x_2, x_3, x_4, x_5) = (168, 170, 193, 187, 201) \rightarrow \bar{x} = 183{,}8$
$(x_1, x_2, x_3, x_4, x_5) = (179, 174, 178, 188, 179) \rightarrow \bar{x} = 179{,}6$
$(x_1, x_2, x_3, x_4, x_5) = (160, 175, 189, 182, 170) \rightarrow \bar{x} = 175{,}2$

$$\vdots \qquad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$$

$$X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_5 \quad \rightarrow \bar{x} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5},$$

i així, podrem imaginar «totes» les mostres de cinc persones que es poden seleccionar amb una població de N habitants.

Si imaginem aquestes dades per columnes, és fàcil observar que els valors de la primera columna són les dades de la variable X_1, que correspon a l'alçada de la primera persona seleccionada en la mostra; així mateix, es pot concloure que són els valors d'una variable aleatòria (presumiblement normal) que té per mitjana el valor μ de la mitjana de l'alçada de tota la població. Aquest mateix raonament es pot fer per a la resta de les columnes, que defineixen, respectivament, les variables X_2, X_3, X_4, X_5.

Així, podrem dir que una mostra aleatòria simple de grandària n és un conjunt de n variables aleatòries $X_1, X_2, X_3, \ldots X_n$ on cada X_i representa el valor observat en la i-èsima extracció i podrà, doncs, agafar qualsevol valor de la població. Totes aquestes variables X_i tenen la mateixa distribució de probabilitat, és a dir, tenen la mateixa funció de probabilitat (en el cas de les discretes) o de densitat (en el cas de les contínues), que anomenarem *distribució de la població*. A més a més, és clar que aquestes variables són *independents*.

Com que les variables que componen la mostra són aleatòries, qualsevol estadístic, calculat amb les dades de la mostra, també és una variable aleatòria. Per tant, un primer pas en la inferència estadística consisteix a analitzar les distribucions de probabilitat dels estadístics per a saber com de fiables són els resultats que obtenim basant-nos-hi. Les distribucions de probabilitat dels estadístics s'anomenen *distribucions en el mostreig.*

8.4. Models de distribució de probabilitat en el mostreig

Al llarg del tema, arribarem a definir les distribucions dels estadístics amb els quals volem abordar la inferència en aquest treball i podrem conèixer el model de distribució que s'ajusta a la mitjana mostral, a la proporció mostral, a la diferència de mitjanes, al quocient de variàncies, etc.

Primerament, cal introduir altres models de distribucions de probabilitat per a afegir-los als que ja coneixem (distribució binomial, de Poisson, normal, etc.), ja que els necessitarem per al nostre propòsit per tractar-se de models que apareixen relacionats amb les distribucions dels estadístics de mostres aleatòries de poblacions normals.

8.4.1. Models de distribució de probabilitat en el mostreig

Khi quadrat

Considerem Z_1, Z_2, ... Z_n un conjunt de variables aleatòries independents que es distribueixen segons el model normal de mitjana 0 i desviació típica 1, és a dir, $Z_1 \rightarrow N(0, 1)$, i definim la variable $X = \sum_{i=i}^{n} Z_i^2$. Direm que la variable X té una distribució khi quadrat amb n graus de llibertat i ho denotarem així $X \rightarrow X_n^2$.

D'aquesta distribució podem afirmar que $E[X] = n$ i que $\text{Var}[X] = 2n$.

En la figura, com a exemple, mostrarem la gràfica de la seua funció de densitat per a $n = 10$ graus de llibertat.

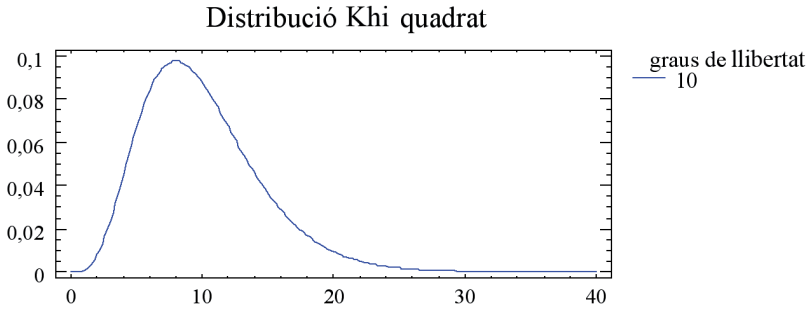


Figura 1

Observem que aquesta funció no és simètrica. Aquest model de distribució apareixerà en les estimacions de la variància de la població, com podrem veure més endavant. A més a més, coneixerem els valors de les seues funcions de distribució $F(x)$ mitjançant taules o programes informàtics d'estadística.

T de Student

Considerem X i Z dues variables aleatòries independents que es distribueixen respectivament com una khi quadrat amb n graus de llibertat i com una variable normal tipificada, és a dir, $X \rightarrow X_n^2$, i $Z_1 \rightarrow N(0, 1)$, i definim la variable $T = \frac{Z}{\sqrt{\frac{X}{n}}}$.

Direm que la variable t es distribueix com una variable t de Student amb n graus de llibertat i ho denotarem així: $T \rightarrow t_n$.

D'aquesta distribució es pot afirmar que $E[X] = 0$.

En la figura següent, com a exemple, mostrarem la funció de densitat de la gràfica per a $n = 10$ graus de llibertat.

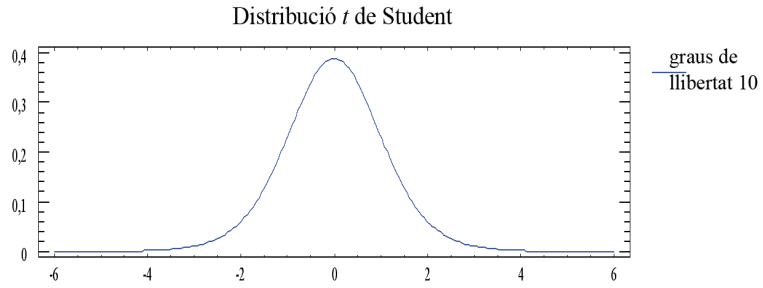


Figura 2

Com es pot advertir en la figura 2, la forma general d'aquesta funció de densitat és similar a la d'una distribució normal; ambdues tenen forma de campana i són simètriques respecte de la mitjana. Igual que amb la normal tipificada, la variable T té una mitjana igual a 0. No obstant això, la seua variància depèn del paràmetre que hem anomenat *graus de llibertat*. La variància de la variable T és superior a 1, però s'apropa a 1 quan n augmenta. De fet, és possible demostrar que la distribució t amb n graus de llibertat tendeix a la distribució normal tipificada quan $n \to \infty$.

En les fórmules que utilitzarem en el capítol següent en el càlcul dels intervals podrem operar $P(T \leq x) = 1 - P(T \leq -x)$ per la simetria d'aquesta funció de densitat.

Aquesta distribució t de Student apareixerà en les distribucions de la mitjana i de la diferència de mitjanes quan la variància poblacional és desconeguda, i coneixerem els valors de les seues funcions de distribució $F(x)$ mitjançant taules o programes informàtics d'estadística.

F de Fisher-Snedecor

Considerem X_1, X_2, dues variables aleatòries independents que es distribueixen segons el model de khi quadrat amb n i m graus de llibertat respectivament, és a dir, $X_1 \to \chi^2_n$, $X_2 \to \chi^2_m$ i definim la variable $F = \frac{\frac{X_1}{n}}{\frac{X_2}{m}}$. Direm que F es distribueix com una variable F de Fisher-Snedecor amb n i m graus de llibertat i ho denotarem així $F \to f_{n,m}$.

En la figura següent es mostra un exemple de la seua funció de densitat per a $n =$ $= 10$ i $m = 5$ graus de llibertat.

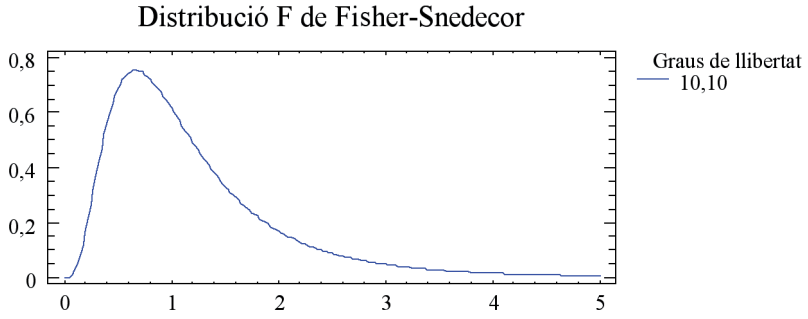


Figura 3

Aquest model de distribució ens apareixerà en les estimacions del quocient de variàncies. Per al càlcul dels valors de la seua funció de distribució cal fer ús de la relació $P(f_{n,m} \leq x) = 1 - P(f_{m,n} \leq \frac{1}{x})$ quan s'empren les taules per a trobar-los.

8.5. Models de distribució de probabilitat d'alguns estadístics

8.5.1. Models per a una mostra

En aquest apartat considerarem $X_1, X_2, ... X_n$ una mostra aleatòria d'una població que es distribueix amb qualsevol model de distribució amb mitjana μ i variància σ^2.

Nota

Al llarg del capítol es parlarà de *distribució de probabilitat de la població, paràmetre de la població*, etc., en lloc de parlar de *distribució de la variable que s'està estudiant sobre tota la població*, *paràmetre relatiu a la variable aleatòria que s'està estudiant sobre tota la població*, etc. Aquest abús de notació es realitza per motius obvis.

Distribució de la mitjana mostral (coneguda σ^2 poblacional)

Suposem que volem estimar el valor del paràmetre μ d'una població. Per a això, agafaríem una mostra de *n* elements i en calcularíem la mitjana mostral.

Considerem la variable aleatòria $\overline{X} = \frac{X_1 + X_2 + ... + ... X_n}{n}$, que assigna a cada mostra de grandària *n* el valor de la seua mitjana.

Aquesta variable aleatòria es diu *mitjana mostral* i la distribució que segueix es diu *distribució mostral de les mitjanes*. Es pot demostrar que:

- La mitjana d'aquesta variable aleatòria $E[\overline{X}] = \mu_{\overline{X}} = \mu$, és a dir, la seua esperança matemàtica és igual a la mitjana de la població.

- La desviació típica de la variable aleatòria $\overline{X}$ és: $\sqrt{\text{Var}\left[\overline{X}\right]} = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

  On σ és la desviació típica poblacional i *n* la grandària de la mostra. D'aquesta propietat es dedueix que les mitjanes mostrals estan més disperses respecte al valor de μ si *n* és menuda, i més properes al valor que volem estimar si augmenta la grandària de la mostra.

- La distribució de la variable aleatòria $\overline{X}$, que anomenem *distribució mostral de les mitjanes*, segueix una distribució normal $\left(\overline{X} \to N\left(\mu_{\overline{X}}, \sigma_{\overline{X}}\right)\right)$ sempre que considerem un mostreig aleatori simple i que la mostra aleatòria tinga una distribució normal. Si no es dóna aquesta darrera exigència, podem aproximar la distribució mostral de les mitjanes per una distribució normal si *n* és gran. A més a més, pel teorema del límit central aquesta aproximació és millor com més gran és *n*. Tanmateix, en cas que la població no seguisca una distribució normal i *n* siga menut, $\overline{X}$ segueix una distrinució *t* de Student amb *n* graus de llibertat. (Ho veurem amb detall en el pròxim apartat.)

Nota

Com que el text és clarament de caràcter docent i introductori, podem considerar que la mostra té una grandària qualificable de *gran* si $n > 30$.

Per altra part, cal recordar que el fet que $\overline{X} \to N\left(\mu_{\overline{X}}, \sigma_{\overline{X}}\right)$ és equivalent a afirmar que la variable $Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ es distribueix com una normal tipificada, és a dir, $Z \to N(0, 2)$.

Com a conseqüència d'aquests resultats podrem tenir:

- Control de les mitjanes mostrals

  En una població de mitjana μ i desviació típica σ, ens disposem a extraure una mostra de grandària *n*. Abans de fer-ho, coneixem la distribució de les mitjanes $\overline{X}$, i podrem esbrinar la probabilitat que la mitjana d'una mostra estiga en un cert interval de valors.

- Control de la suma dels valors d'una mostra

  Podrem calcular la probabilitat que la suma dels valors d'una mostra estiga dins d'un interval, ja que $\sum_{i=1}^{n} X_i = n\overline{X}$ i la distribució d'aquesta darrera és coneguda.

- Inferència la mitjana de la població a partir de la mitjana d'una mostra

  Aquesta és l'aplicació més important i la que ens havíem plantejat des del començament del tema. En aquesta unitat parlarem de l'estimació puntual i en la propera unitat parlarem de l'estimació per intervals, amb cert grau de certesa que denominarem *nivell de confiança* i que indicarà, en termes de probabilitat, la certesa dels nostres resultats.

Exemple 1

Els paquets de farina envasats per una certa màquina tenen $\mu = 500$ g i $\sigma = 35$ g, i s'embalen en caixes de 100 unitats.

a. Calculem la probabilitat que la mitjana dels pesos dels paquets d'una caixa siga inferior a 495 g.
b. Calculem la probabilitat que una caixa de 100 paquets pese més de 51 kg.

Comencem per identificar la totalitat de paquets de farina envasats per la màquina com la població de mitjana $\mu = 500$ g i desviació típica $\sigma = 35$ g. Cada caixa serà una mostra de grandària $n = 100$.

Les mitjanes dels pesos dels paquets d'una caixa es distribueixen com una variable normal de mitjana $\mu = 500$ g i desviació típica $\frac{\sigma}{\sqrt{n}} = \frac{35}{\sqrt{100}} = 3{,}5$. Així doncs, $\overline{X} \rightarrow N(500; 3, 5)$.

- Per a calcular la probabilitat que la mitjana dels pesos dels paquets d'una caixa siga inferior a 495 g, transformarem la variable $\overline{X}$, mitjançant una tipificació, i acudirem a un paquet informàtic per a trobar el valor de la variable en la funció de distribució de la variable normal tipificada Z:

$$P(\overline{X} < 495) = P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} < \frac{495 - 500}{3{,}5}\right) = P(Z < -1{,}43) = \phi(-1{,}43) = 1 - \phi(1{,}43) =$$

$$= 1 - 0{,}9236 = 0{,}0764.$$

- Per a calcular la probabilitat que una caixa de 100 paquets pese més de 51 kg, considerarem la variable $\sum_{i=1}^{n} X_i$, que també es distribueix com una normal però de mitjana $n\mu$ i de desviació típica $n\frac{\sigma}{\sqrt{n}} = \sigma\sqrt{n}$. Per tant, $\sum_{i=1}^{n} X_i \to N(50.000;\ 350)$.

$$P\left(\sum_{i=1}^{n} X_i > 51000\right) = P\left(\frac{\sum_{i=1}^{n} X_i - 50000}{350} > \frac{51000 - 50000}{350}\right) = P(Z > 2{,}86) =$$

$$= 1 - P(Z \leq 2{,}86) = 1 - \phi(2.86) = 1 - 0{,}9979 = 0{,}0021.$$

Distribució de la mitjana mostral (si no és coneguda σ^2 poblacional)

Quan no es coneix el valor de la σ^2 poblacional, cal emprar una altra distribució per a abordar un estadístic que explique el comportament de la mitjana mostral, sense que aquest paràmetre aparega. Així, s'utilitza l'expressió següent, que es defineix com a quocient d'una variable normal tipificada i l'arrel quadrada d'una variable khi quadrat amb $n - 1$ graus de llibertat.

Si denotem per $\overline{X}$ la variable aleatòria mitjana mostral de grandària n, extreta d'una població normal de mitjana μ i variància σ^2, i per $S = S^2_{X_{n-1}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ la quasi variància mostral, es pot demostrar que la variable aleatòria $T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$ es distribueix segons una distribució t de Student amb $n - 1$ graus de llibertat. És a dir, $T \to t_{n-1}$.

Aquest resultat és més general que l'anterior, ja que no necessitem conèixer el valor de la variància poblacional σ^2. Tanmateix, és més restrictiu perquè implica el supòsit d'una població normal. Malgrat tot, aquesta restricció no és tan severa com semblaria. Els estudis han demostrat que la distribució de la variable aleatòria $T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$ s'apropa molt a una distribució t de Student, encara que considerem mostres que no provenen d'una població normal. En la pràctica serà suficient poder assegurar-nos que la població de la qual fem el mostreig té una distribució en forma de campana i no massa esbiaixada.

Exemple 2

Un fabricant de fusibles assegura que amb una sobrecàrrega del 20% els fusibles es fondrien en una mitjana de 12,40 minuts de mitjana. Per a comprovar aquesta afirmació agafem 20 fusibles d'aquest material i els sotmetem a una sobrecàrrega del 20% i calculem els temps que tardaran a fondre's. Hi obtenim una mitjana de 10,63 minuts i una quasidesviació típica de 2,48 minuts. Suposem que les dades provenen d'una distribució normal, podem asseverar o refutar la informació del fabricant?

Calcularem el valor de la variable T amb les nostres dades poblacionals i mostrals:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{10{,}63 - 12{,}40}{2{,}48/\sqrt{20}} = -3{,}19,$$

el qual serà un valor de la distribució t de Student amb 19 graus de llibertat. Si l'afirmació és certa, seria lògic que el nostre valor estiguera, per exemple, dins del 90% dels valors centrals d'aquesta distribució. Així calcularem els valors a i b que compleixen: $P(t \leq a) = 0{,}05$ i $P(t \geq b) = 0{,}05$, és a dir $P(t \leq b) = 0{,}95$. Si el valor $t = -3{,}19$ està fora d'aquest interval, direm que la nostra prova no corrobora l'afirmació del fabricant.

Amb l'ajuda d'un programa d'estadística calculem $a = -1{,}73$ i, per simetria, $b = +1{,}73$, per la qual cosa el valor que hem obtingut està fora d'aquest interval i ens porta a refutar la informació prèvia que tenim del valor de μ.

Distribució d'una proporció

En una determinada població, la proporció d'individus que posseeixen una característica determinada la denotarem per p. Per a esbrinar aquest valor, es consideren totes les possibles mostres de grandària n que es poden extraure d'aquesta població. En cada una de les mostres hi haurà una proporció p_1, p_2, ... d'individus de la població que posseeixen aquesta característica. Estudiem com es distribueixen tots els possibles valors de $\mathbf{p}$.

Si denotem per X el nombre d'individus de la mostra que tenen una determinada característica, podem afirmar pels coneixements de temes anteriors que $X \rightarrow B(n, \mathbf{p})$; també vam veure que si es compleix que $np \geq 5$ i que $nq \geq 5$ (on òbviament $q = 1 - p$), podem transformar una variable binomial en una de normal i assegurar que $X \rightarrow N(np, \sqrt{npq})$.

Si denotem per $p = \dfrac{\text{nombre d'individus amb la característica determinada}}{\text{nombre d'individus de la mostra}} = \dfrac{X}{n}$, la distribució de p, també anomenada *distribució mostral de la proporció*, serà la mateixa que la de X, però amb els paràmetres de mitjana i desviació típica dividits per n. Per tant:

$$\mathrm{p} \to \mathrm{N}\left(\frac{np}{n}; \frac{\sqrt{npq}}{n}\right) = \mathrm{N}\left(p; \sqrt{\frac{pq}{n}}\right).$$

Aquesta aproximació és millor com més gran siga la mostra i més propera estiga p de 0,5. Així, si les condicions inicials ($np \geq 5$ i $nq \geq 5$) es compleixen, podem considerar que tenim una bona aproximació. Si aquest no fóra el cas, caldria augmentar la grandària de la mostra.

Exemple 3

Suposem que el percentatge de famílies a la Comunitat Valenciana que tenen un fill únic és del 20%. Considerem una mostra de 1.000 famílies i volem esbrinar quina és la probabilitat que almenys el 21% d'aquestes famílies tinga un fill únic.

En aquest cas $n = 1.000$ i $p = 0{,}20$, per tant la variable aleatòria que indica les proporcions mostrals p té un distribució normal:

$$\mathrm{p} \to \mathrm{N}\left(\frac{np}{n}; \frac{\sqrt{npq}}{n}\right) = \mathrm{N}\left(p; \sqrt{\frac{pq}{n}}\right) = N\left(0{,}2; \sqrt{\frac{0{,}2 \cdot 0{,}8}{1000}}\right) = N(0{,}2;\ 0{,}0126).$$

Amb aquesta distribució, la probabilitat que ens plantegem serà:

$$P(\mathrm{p} \geq 0{,}21) = P\left(\frac{\mathrm{p} - \mu}{\sigma} \geq \frac{0{,}21 - 20}{0{,}0126}\right) = P(Z \geq 0{,}79) = 1 - P(Z \leq 0{,}79) = 0{,}2148.$$

I podem considerar que és una bona aproximació, ja que $p \cdot n = 0{,}2 \cdot 1.000 = 20 > 5$ i que $q \cdot n = 0{,}8 \cdot 1.000 = 800 > 5$.

Distribució de la variància

No podem trobar un model de distribució que ens permeta conèixer directament la distribució de l'estadístic *variància mostral*, però sí que podem conèixer la distribució d'aquesta variable que definirem, que permetrà conèixer el valor de la variància de la població, basant-nos en la variància de la mostra.

Si denotem per $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ la quasivariància d'una mostra aleatòria de grandària n, extreta d'una població normal de variància σ^2, es pot demostrar que la variable aleatòria $X = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2}$ es distribueix com una variable khi quadrat amb $n - 1$ graus de llibertat, és a dir, $X \rightarrow X^2_{n-1}$.

Recordem dels apartats anteriors que la variable khi quadrat, l'havíem definit com la suma de variables normals tipificades elevades al quadrat $X = \sum_{i=i}^{n} Z_i^2$ i aquesta estructura, podem trobar-la en l'expressió de la nostra variable.

Exemple 4

Imaginem que una màquina envasadora de fruites seques té una certa dispersió respecte al pes que marca l'etiqueta. Donem per correcte el funcionament de la màquina si la variància dels pesos no excedeix d'1,26 g. Quan acaba el dia, per a fer un seguiment del procés agafem 20 peces envasades i calculem la seua variància en els pesos. Si aquesta excedeix de 2 g, considerem que cal revisar la màquina i avisem el tècnic. Quina és la probabilitat d'haver avisat el tècnic i que la màquina funcione dins dels paràmetres estipulats?

Per a obtenir el valor de la variable se substitueixen els valors coneguts de la població i de l'estadístic en l'expressió de la variable X:

$$X = \frac{(n-1)S^2}{\sigma^2} = \frac{19 \cdot 2}{1{,}26} = 30{,}16.$$

Com que aquest és un valor d'una variable X^2 amb 19 graus de llibertat, podem veure en la figura següent la seua funció de densitat i observar que el valor 30,16 està a la part dreta de la distribució. A més a més, amb les taules es pot calcular que $P(X \leq 30{,}16) = 0{,}95$. Així, la probabilitat d'aturar la màquina en funció del valor de la variància de la mostra i que aquesta sí que funcione correctament és tan sols del 5%.

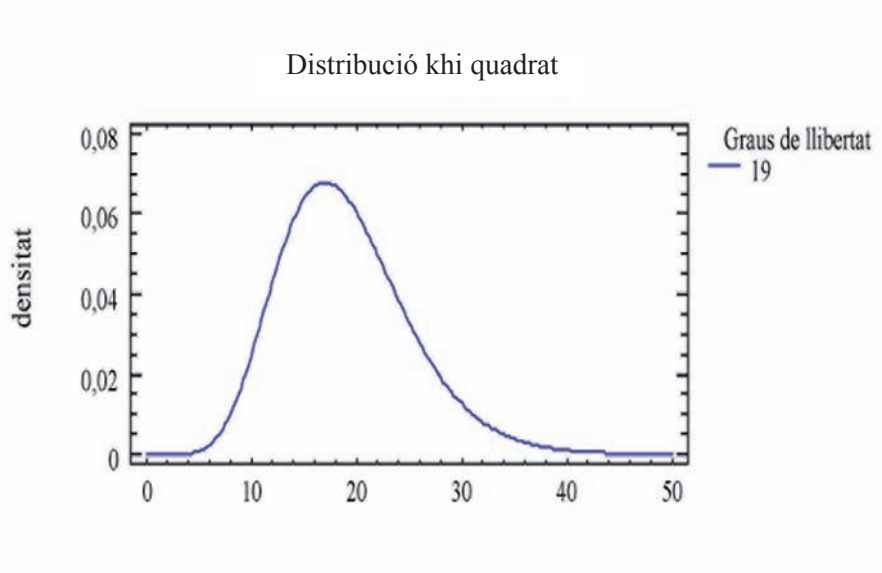


Figura 4

8.5.2. Models per a dues mostres

Per a començar, cal diferenciar dues possibles circumstàncies pel que fa a les composicions de les mostres. Així, cal saber si estem en el cas de mostres independents de dues poblacions (alumnat de dos grups diferents; treballadors de dos seccions diferents; homes enfront de dones per a estudiar algun comportament lligat o no amb el gènere, etc.) o si es tracta de mostres relacionades o aparellades (si estudiem el desgast en les rodes dreta i esquerra d'un vehicle; la duració d'un material de sola d'un calcer esportiu per al qual s'ha dissenyat un model, i es proporciona a cada individu una sabata de cada peu amb cada material; les mesures de la pressió arterial preses en cada malalt abans i després de seguir un tractament farmacològic, etc.).

Suposem ara que volem comparar dues variables de dues poblacions. Necessitem, doncs, triar dues mostres, una de cadascuna. A més, suposem que cada mostra és independent. En cas de considerar-les relacionades o emparellades s'especificarà. En els apartats següents, denotarem per $X_1, X_2, \ldots X_n$ una mostra aleatòria de grandària *n* extreta d'una població que es distribueix amb qualsevol model de distribució amb mitjana μ_x i variància σ_X^2; i Y_1, Y_2, ..., Y_m una mostra aleatòria de grandària *m* extreta d'una altra població que es distribueix amb qualsevol model de distribució amb mitjana μ_Y i variància σ_Y^2.

Diferència de mitjanes

Sovint comparem les mitjanes de dues poblacions per veure si el rendiment mitjà és millor en una població que en l'altra. Així, podríem comparar els resultats d'una prova d'estadística en un grup d'estudiants del torn del matí i comparar-la amb els del grup de la vesprada. Per a això, tal com estem treballant al llarg d'aquest

tema, caldrà triar una mostra representativa de les proves de l'alumnat de cadascun dels grups i traure la mitjana mostral de cadascuna. Imaginem que una mitjana és unes dècimes més alta que l'altra. Amb els plantejaments propers podrem decidir si aquesta diferència entre les mitjanes és realment significativa o és causada per l'atzar.

A continuació, presentem les distribucions de tres variables que ens permetran fer aquesta comparació en tres circumstàncies que cal considerar a l'hora de comparar les mitjanes:

- Mostres independents amb variàncies poblacionals conegudes
- Mostres independents amb variàncies poblacionals no conegudes, però iguals
- Mostres dependents de dades relacionades ($n = m$ necessàriament).

Distribució de la diferència de les mitjanes (si coneixem σ^2_X i σ^2_Y)

Considerem les variables aleatòries $\overline{X} = \frac{X_1 + X_2 + ... + ... X_n}{n}$ i $\overline{Y} = \frac{Y_1 + Y_2 + ... + Y_m}{m}$, que assignen a cada mostra el valor de la seua mitjana, i definim a partir d'aquestes la variable aleatòria $\overline{X} - \overline{Y}$. Es pot demostrar que aquesta variable es distribueix com una variable normal que té per mitjana la resta de les mitjanes mostrals $\mu_{\bar{x}-\bar{y}} = \mu_{\bar{x}} - \mu_{\bar{y}} = \mu_x - \mu_y$ i per variància, la suma de les variàncies de les mitjanes mostrals $\sigma^2_{\bar{X}-\bar{Y}} = \sigma^2_{\bar{X}} + \sigma^2_{\bar{Y}} = \frac{\sigma^2_X}{n} + \frac{\sigma^2_Y}{m}$. Així:

$$\overline{X} - \overline{Y} \to N\left(\mu_X - \mu_Y, \sqrt{\frac{\sigma^2_X}{n} + \frac{\sigma^2_Y}{m}}\right).$$

Una expressió equivalent a l'anterior, que s'obté en tipificar:

$$\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2_X}{n} + \frac{\sigma^2_Y}{m}}} \to N(0, 1).$$

Es pot demostrar que aquest resultat és vàlid si les poblacions de les quals hem extret les mostres es distribueixen com una variable normal. No obstant això, el resultat també es compleix quan treballem amb mostres suficientment grans que ens permeten aplicar el teorema del límit central i aproximar σ^2_X i σ^2_Y per les respectives quasivariàncies. En els exemples d'aquest manual considerarem adient l'aproximació quan n i m són superiors a 30.

Exemple 5

Coneixem que la duració mitjana de funcionament dels frigorífics de la marca A és de 18 anys i els de la marca B, de 16 anys; també coneixem que les desviacions típiques són 3 i 5 anys respectivament. Seleccionem aleatòriament 75 frigorífics de la marca A i 50 de la marca B i anotem les seues duracions. Calculem la probabilitat que la mitjana de la mostra dels frigorífics de la marca A supere en més d'un any la mitjana dels de la marca B.

Les dades donades per l'enunciat són:

Marca A:	$n = 75$	$\mu_x = 18$	$\sigma_X = 3$
Marca B:	$m = 50$	$\mu_y = 16$	$\sigma_Y = 5$

Podem veure la distribució de la variable $\overline{X} - \overline{Y}$:

$$\overline{X} - \overline{Y} \to N\left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right) = N\left(18 - 16, \sqrt{\frac{9}{75} + \frac{25}{50}}\right) = N(2; 0{,}787).$$

Per a calcular la probabilitat plantejada haurem de tipificar i trobar el valor corresponent de la probabilitat en la funció de distribució de la variable *Z*, normal tipificada.

$$P\left(\overline{X} - \overline{Y} > 1\right) = P\left(\frac{\overline{X} - \overline{Y} - 2}{0{,}787} > \frac{1-2}{0{,}787}\right) = P\left(Z > -1{,}27\right) = 1 - P\left(Z \le -1{,}27\right) =$$

$$= 1 - \phi\left(-1{,}27\right) = 1 - \left(1 - \phi\left(1{,}27\right)\right) = \phi\left(1{,}27\right) = 0{,}8980.$$

Distribució de la diferència de les mitjanes (si no coneixem σ_Y^2 i σ_Y^2)

Quan desconeixem les variàncies poblacionals no podem aplicar la distribució de la variable que hem presentat en l'apartat anterior. No obstant això, podem emprar-ne una altra sempre que puguem assumir que les variàncies són iguals, encara que siguen desconegudes.

Així doncs, caldrà fer primerament un treball d'inferència sobre el quocient de les variàncies de la població amb l'objecte de saber si les variàncies poden considerar-se iguals. Aquest treball d'inferència es tractarà més endavant. Així mateix, tan sols quan siga possible inferir l'esmentada igualtat de variàncies podrem aplicar el resultat que tot seguit desenvolupem.

Si denotem per

$$S^2_{X_{n-1}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

la quasivariància de la mostra aleatòria de grandària n extreta de la primera població i per:

$$S^2_{Y_{m-1}} = \frac{\sum_{i=1}^{m}(y_i - \bar{y})^2}{m-1}$$

la quasivariància de la mostra de grandària m de la segona població, es pot demostrar que la variable aleatòria:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{(n-1)S^2_{x_{n-1}} + (m-1)S^2_{Y_{m-1}}}{n+m-2}}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

es distribueix segons el model d'una variable t de Student amb $n + m - 2$ graus de llibertat. És a dir, $T \rightarrow t_{n+m-2}$.

Nota

Aplicarem aquest resultat quan n i m siguen inferiors a 30 i les mostres, independents.

Exemple 6

Volem comprovar l'eficàcia d'una màquina classificadora de fruita pel calibre. Perquè una fruita siga classificada de categoria A cal que el seu diàmetre faça 7,25 cm de mitjana, i per a classificar-les de categoria B, 5,5 cm. Suposem que les desviacions típiques d'aquestes mesures dels diàmetres són iguals en totes dues categories.

Hem triat 12 fruites que la màquina ha calibrat com a categoria A, amb una desviació típica dels seus pesos de 0,98 cm, i 15 fruites de les classificades com a categoria B, amb una desviació típica d'1 cm. Calculem la probabilitat que les mitjanes del calibre de les fruites d'aquestes mostres tinguen una diferència superior a 1,25 cm de diàmetre.

Si entenem que els valors de l'enunciat corresponen a les quasidesviacions típiques, les dades donades en l'enunciat són les següents:

Categoria A:	$n = 12$	$\mu_x = 7{,}25$	$S^2_{X_{n-1}} = 0{,}98^2 = 0{,}9604$
Categoria B:	$m = 15$	$\mu_Y = 5{,}5$	$S^2_{Y_{m-1}} = 1^2 = 1.$

Podem veure la distribució de la variable T si substituïm els valors de la mostra i la diferència entre les mitjanes de les poblacions corresponents. Així:

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{(n-1)S^2_{x_{n-1}} + (m-1)S^2_{Y_{m-1}}}{n+m-2}}\sqrt{\frac{1}{n}+\frac{1}{m}}} = \frac{1{,}25 - (7{,}25 - 5{,}5)}{\sqrt{\frac{11 \cdot 0{,}9604 + 14 \cdot 1}{25}}\sqrt{\frac{1}{12}+\frac{1}{15}}} =$$

$$= \frac{-0{,}5}{0{,}9912 \cdot 0{,}3873} = \frac{-0{,}5}{0{,}3839} = -1{,}3024.$$

La probabilitat que la diferència de les mitjanes siga superior a 1,25 equival a calcular la probabilitat que aquest estadístic tinga valors superiors a –1,3024. Així calcularem, amb l'ajuda d'un programa o de les taules, aquesta probabilitat en una distribució t de Student amb 25 graus de llibertat:

$$P(\overline{X} - \overline{Y} > 1{,}25) = P(T > -1{,}3024) = 0{,}8977.$$

Distribució de la diferència de mitjanes amb mostres relacionades

En alguns casos pot ser interessant comparar les mitjanes de dues poblacions que estan relacionades, com per exemple la mitjana dels pesos d'uns individus abans i després de fer un tractament per a aprimar-se, les diferències d'uns indicadors de qualitat abans i després de fer una millora en els materials o el procés de producció, etc.

En el disseny de l'experiment sovint és interessant considerar aquesta relació. Per exemple, si volem conèixer la diferència de la qualitat d'un material per a pneumàtics, és un bon experiment posar en la roda dreta el material A i en l'esquerra, el material B, perquè una mostra aleatòria de vehicles forme part de l'experiment. Així, estarem assegurant-nos que els materials estan sotmesos a les mateixes condicions de conducció, com ara temperatura, terreny, forma de conducció, etc.

Són, doncs, molts, els experiments que ens portaran a fer aquesta comparació de mitjanes amb mostres que podem anomenar *relacionades* o *aparellades*.

Per a dur a terme aquesta distribució, crearem una variable, que podem anomenar *diferència* i que denotarem per D, que pren com a valors la diferència de cada parell. Així, calcularem $D_i = X_i - Y_i$ i amb aquestes dades, calcularem $S_{D_{n-1}}$, que serà la quasidesviació típica d'aquests valors.

Si definim la variable:

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\frac{S_{D_{n-1}}}{\sqrt{n}}},$$

es pot demostrar que té una distribució que segueix el model d'una variable t de Student amb $n - 1$ graus de llibertat. És a dir, $T \rightarrow t_{n-1}$.

Exemple 7

Volem demostrar que no hi ha diferència entre els rendiments de treball de dos departaments d'una empresa. Per a comprovar-ho vam dissenyar 10 treballs iguals que vam encarregar a tots dos departaments i en cadascun vam anotar el grau d'assoliment en una setmana.

Amb les dades de les nostres mostres, vam calcular $S_{D_{n-1}} = 0{,}1792$. Nosaltres creiem que no hi ha diferència en els rendiments mitjans entre tots dos departaments, així $\mu_x - \mu_Y = 0$. Calculem la probabilitat que, agafant mostres de 10 treballs, la diferència entre les mitjanes mostrals siga superior a 0,5.

Per a calcular aquesta probabilitat, calcularem el valor d'aquesta variable T per a les nostres dades, i calcularem les probabilitats mitjançant programes estadístics, atès que coneixem que es distribueix com una variable t de Student amb 9 graus de llibertat:

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_Y)}{\frac{S_{D_{n-1}}}{\sqrt{n}}} = \frac{0{,}5 - 0}{\frac{0{,}1792}{\sqrt{10}}} = \frac{0{,}5}{0{,}0567} = 8{,}8183.$$

Per l'expressió d'aquesta variable, $P(\overline{X} - \overline{Y}) = P(T \geq 8{,}8183) = 0$, com podem observar també en la figura següent.

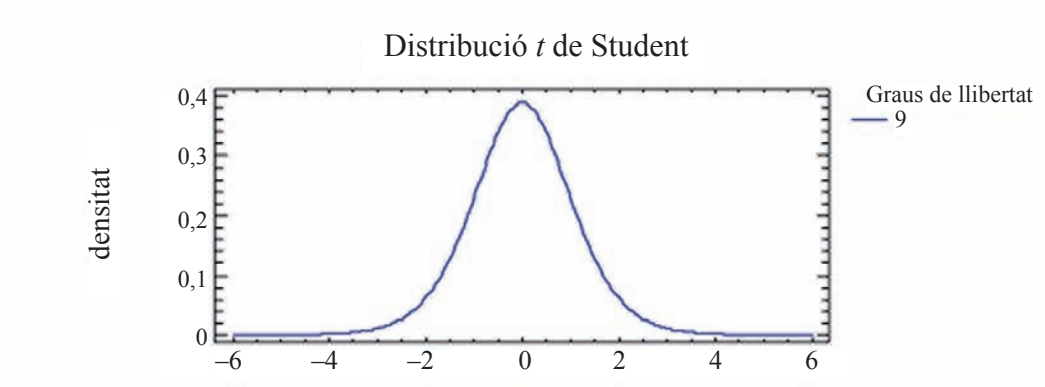


Figura 5

Distribució de la diferència de proporcions

En alguns casos, ens podrem plantejar comparar les proporcions d'individus de dues poblacions amb una determinada característica. Ens pot interessar comparar el percentatge de productes defectuosos en dos processos de producció, el percentatge de votants d'un determinat partit polític en dues eleccions consecutives, el percentatge d'afectats de certa malaltia en dues comunitats, etc. En tots aquests exemples, calculem la proporció dels individus en cadascuna de les mostres i volem inferir si les diferències entre els valors que s'obtenen en totes dues mostres són significatives i estan al voltant d'un determinat valor o si tan sols estan causades per un procés d'atzar per tractar-se de mostres triades aleatòriament que poden diferir mínimament en els valors concrets.

En aquest apartat denotarem per $X_1, X_2, ..., X_n$ una mostra aleatòria de grandària n extreta d'una població en la qual la proporció d'individus que té una determinada característica serà denotada per p_x. La proporció que obtenim amb aquests valors de la mostra, la denotarem per $\mathbf{p_x}$.

Igualment, referirem per $Y_1, Y_2, ..., Y_m$ la mostra aleatòria de grandària m extreta d'una altra població que té una proporció que denotarem per p_x. El valor de la proporció calculat amb les dades d'aquesta mostra, el denotarem per $\mathbf{p_y}$.

La distribució de $\mathbf{p_x}$, també anomenada *distribució mostral de la proporció*, ja la vam veure anteriorment en analitzar el cas d'una mostra, i correspon a una distribució normal amb els paràmetres següents:

$$p_X \to N\left(p_x; \sqrt{\frac{p_X(1-p_X)}{n}}\right)$$ i si la transformem per tipificar-la, $$\frac{p_X - p_x}{\sqrt{\frac{p_X(1-p_X)}{n}}} \to N(0, 1).$$

Paral·lelament obtenim una expressió semblant de la distribució de $\mathbf{p_y}$ i de la seua tipificació:

$$\frac{p_Y - p_Y}{\sqrt{\frac{p_Y(1-p_Y)}{m}}} \to N(0, 1).$$

Si ens plantegem la distribució de la seua diferència $\mathbf{p_x} - \mathbf{p_y}$ és la diferència de dues distribucions normals tipificades, per la qual cosa la seua distribució serà una normal que té per mitjana la resta de les mitjanes i per variància, la suma de les variàncies de cadascuna. Així, coneixem la distribució de la variable que pren com a valor la diferència de les proporcions de les mostres extretes de totes dues poblacions.

Definim la variable:

$$Z = \frac{(p_x - p_Y) - (\mathrm{p_X} - \mathrm{p_Y})}{\sqrt{\frac{p_x(1-p_x)}{n} + \frac{p_Y(1-p_Y)}{m}}} \to N(0,\ 1),$$

que es pot demostrar que es distribueix segons una distribució normal tipificada de mitjana 0 i variància 1, quan la grandària de les mostres és gran, la qual cosa permet transformar una variable binomial en una de normal.

Exemple 8

Per comparar l'eficàcia de dos tractaments de plagues, mesurem en les mostres la proporció d'arbres sans de diferents camps de cultius que han estat tractats amb cadascun dels productes.

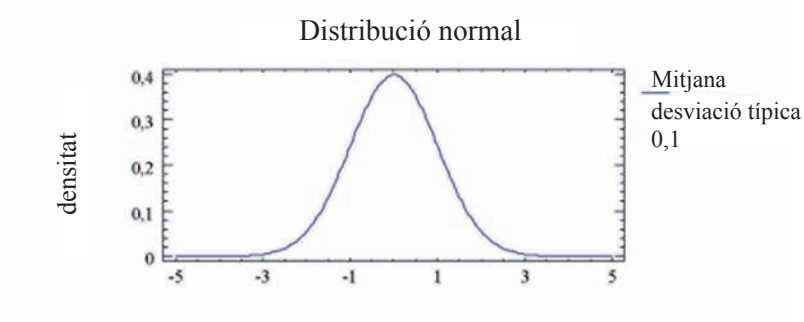
Considerarem que el producte A és millor que el producte B, ja que la diferència de la proporció és del 25%. Podem veure que el més probable és que, amb aquestes dades, obtinguem unes proporcions mostrals que també mostren una diferència del 25% més o menys.

Podem veure que, si observem l'expressió de la variable Z que hem definit, en aquest supòsit que hem fet de diferències iguals, el numerador s'anul·la, independentment del valor de les proporcions poblacionals i mostrals, així com de la grandària de les mostres:

$$(\mathrm{p_X} - \mathrm{p_Y}) \approx (p_x - p_Y) \to (\mathrm{p_X} - \mathrm{p_Y}) - (p_x - p_Y) \approx 0$$

$$Z = Z = \frac{(p_x - p_Y) - (\mathrm{p_X} - \mathrm{p_Y})}{\sqrt{\frac{p_x(1-p_x)}{n} + \frac{p_Y(1-p_Y)}{m}}} = \frac{0}{\sqrt{\frac{p_x(1-p_x)}{n} + \frac{p_Y(1-p_Y)}{m}}} = 0.$$

Per això, si fem el dibuix de la funció de densitat d'aquesta variable normal tipificada, els valors centrals corresponen als valors de la variable propers a 0, que com podem veure, corresponen als valors mostrals més propers a valors poblacionals.



Distribució del quocient de variàncies

En algunes situacions necessitem comparar les variàncies de dues poblacions. Per exemple, aquest procés és necessari per a calcular l'estimació de la diferència de mitjanes de dues poblacions, on cal comprovar que les variàncies poblacionals poden ser desconegudes però iguals.

Considerarem $X_1, X_2, ..., X_n$ una mostra aleatòria de grandària n extreta d'una població amb variància σ^2_X, i $Y_1, Y_2, ..., Y_m$ una mostra aleatòria de grandària m extreta d'una altra població amb variància σ^2_Y. Les quasivariàncies de cada una de les mostres, les denotem per $S^2_{X_{n-1}}$ i per $S^2_{Y_{n-1}}$, respectivament.

Per a comparar les variàncies de totes dues poblacions, cal estudiar la distribució de la variable següent, que és l'estadístic que ens permet comparar per quocient, ja que d'aquesta expressió podrem obtenir una distribució coneguda, que ja vam veure en l'apartat anterior. Recordem que el model de la variable F de Fisher-Snedecor és un model que s'ajusta al quocient de dues variables que presenten una distribució d'una variable khi quadrat.

Per això, definim la variable aleatòria:

$$F = \frac{\dfrac{(n-1)S^2_{X_{n-1}}}{\sigma^2_X \cdot n}}{\dfrac{(m-1)S^2_{Y_{n-1}}}{\sigma^2_Y \cdot m}},$$

que podem demostrar que té una distribució, que és la de la variable F de Fisher-Snedecor amb $n - 1$ i $m - 1$ graus de llibertat, respectivament.

Recordem que aquest model, el vam definir com una distribució que és el quocient de dues variables que es distribueixen com a khi quadrat de n i m graus de llibertat.

Exemple 9

Podem utilitzar com a exemple les dades de l'exemple 6, on per a comparar les mitjanes, cal pressuposar que les variàncies són iguals.

Recordem que volíem comprovar l'eficàcia d'una màquina classificadora de fruita pel calibre. Per què una fruita siga classificada de categoria A, cal que el seu diàmetre mesure 7,25 cm de mitjana, i per a classificar-les de categoria B, 5,5 cm.

Hem triat 12 fruites que ha calibrat la màquina com a categoria A i la quasidesviació típica dels seus pesos de 0,98 cm, i hem triat 15 fruites de les classificades com a categoria B i hem calculat la seua quasidesviació típica d'1 cm.

Les dades que es donen en l'enunciat són les següents: $n = 12$ i $m = 15$.

Si suposem que les variàncies de totes dues poblacions són iguals, podrem calcular la probabilitat que si agafem dues mostres qualssevol de 12 i 15 fruites de cadascuna de les categories, la diferència entre les seues quasivariàncies mostrals siga superior a un 10%. Notem que en aquest plantejament la comparació entre les variàncies, cal fer-la en termes relatius, ja que el model de distribució s'ajusta a un quocient entre les expressions que desenvolupem a continuació. Així, en el càlcul podrem substituir:

$$\frac{S^2_{X_{n-1}}}{S^2_{Y_{n-1}}} = 1{,}10.$$

Per a això, calcularem el valor de l'estadístic que hem definit com a variable F amb les dades de l'enunciat:

$$F = \frac{\dfrac{(n-1)S^2_{X_{n-1}}}{\sigma^2_X \cdot n}}{\dfrac{(m-1)S^2_{Y_{n-1}}}{\sigma^2_Y \cdot m}} = \frac{\dfrac{(11)S^2_{X_{n-1}}}{\sigma^2_X \cdot 12}}{\dfrac{(14)S^2_{Y_{n-1}}}{\sigma^2_Y \cdot 15}} = \frac{\sigma^2_Y \cdot 15 \cdot 11 \cdot S^2_{X_{n-1}}}{\sigma^2_X \cdot 12 \cdot 14 \cdot S^2_{Y_{n-1}}} = \frac{15.11.1{,}10}{12 \cdot 14 \cdot 1} = \frac{181{,}5}{168} = 1{,}0804.$$

Si hem considerat que $\sigma^2_X = \sigma^2_Y$, podem substituir $\dfrac{\sigma^2_Y}{\sigma^2_X} = 1$.

Per a calcular les probabilitats considerem que aquesta variable *F* es distribueix com una *F* de Fisher-Snedecor amb 11 i 14 graus de llibertat. Aquests valors, els podrem calcular amb l'ajuda dels programes estadístics o en les taules.

Ara calcularem la probabilitat que $P\left(\dfrac{S^2_{X_{n-1}}}{S^2_{Y_{n-1}}} > 1{,}10\right) = P\left(f_{11,14} > 1{,}0804\right) = 0{,}4379.$

Podem veure que aquest percentatge del 10% serà poc adient per a agafar-lo com a indicador, ja que encara que les variàncies poblacionals siguen iguals, quasi la meitat (43,79%) de les mostres triades aleatòriament tindran més d'aquest percentatge de diferència, segons els nostres càlculs.

Nota

Cal fixar-se que en aquest apartat que ara acaba, els exemples no aborden cap problema d'inferència tal com havíem plantejat a l'inici del tema, on ja vam definir que l'objectiu és poder esbrinar el valor d'un paràmetre poblacional desconegut θ, basant-nos en un paràmetre mostral *T*, que calculem amb les dades reals dels valors d'una mostra aleatòria.

Els exemples que hem vist tan sols pretenen demostrar que les variables que hem definit en cada apartat tenen una distribució determinada que ens permet plantejar-nos càlculs de probabilitats i per a conèixer les seues funcions de densitat, la simetria d'aquesta, etc.

8.6. Estimació puntual

Al llarg del tema, hem comentat que l'objectiu de la Inferència Estadística és esbrinar el valor d'un paràmetre poblacional θ, basant-nos en un estadístic T que calculem amb les dades d'una mostra.

Pel desenvolupament dels apartats anteriors, ja sabem que $\bar{x}$ no té exactament el valor de μ, ni **p** coincideix exactament amb el valor de *p*, etc. Podríem dir el mateix de les altres mesures estadístiques, que podem calcular amb les dades de la mostra per a obtenir el valor corresponent de la població.

Aquest procediment és el que definim com a *estimació puntual.* Afirmarem que el valor concret d'un estadístic T, serà aquell que s'aproparà al valor que volem esbrinar de la població θ, però no podem controlar el grau de fiabilitat de la nostra estimació.

Ara bé, el que sí que podem és estudiar una sèrie de propietats que seria desitjable que compliren les distribucions dels estadístics, ja que aquestes són conegudes. Així, entre els diferents estadístics T_1, T_2, ...T_n que ens podem plantejar per a esbrinar un determinat paràmetre θ, triarem el més idoni.

Per exemple, si volem calcular l'alçada de la població que ens havíem plantejat en el primer exemple del tema, podrem triar mostres de cinc individus i plantejar-nos diferents alternatives:

- calcular la mitjana de cada mostra $\overline{X}$.
- calcular la semisuma dels valors més gran i més menut de la mostra $\frac{X_{\text{màx.}} + X_{\text{mín.}}}{2}$.
- donar com a estimació el valor central de la mostra, etc.

Qualsevol d'aquestes estratègies podria ser un estimador de la mitjana μ, ara bé, és evident que, a colp d'ull, no ens semblen totes igual de fiables.

Un altre problema clàssic és decidir entre la variància o la quasivariància mostrals com a dos estimadors de la variància de la població.

En aquest apartat, estudiarem tres propietats que són, en certa manera, una mesura de la «qualitat» d'un estadístic com a estimador. Aquestes són el biaix, la consistència i l'eficiència.

8.6.1. Biaix

Direm que un estimador T és un estimador centrat o no esbiaixat de θ si $E[T] = \theta$. És a dir, la distribució d'aquest estimador està centrada en el paràmetre que volem estimar.

8.6.2. Consistència

Direm que un estimador T és consistent de θ si $\lim_{n\to\infty} E[T] = \theta$ i $\lim_{n\to\infty} \text{Var}[T] = 0$.

És a dir, que quan augmenta la grandària de la mostra, l'estimació millora perquè els valors de l'estadístic estan centrats en θ i disminueix gradualment la dispersió.

8.6.3. Eficiència

Si considerem dos estimadors T_1 i T_2 del paràmetre θ, direm que T_1 és més eficient que T_2 si $\text{Var}[T_1] < \text{Var}[T_2]$, és a dir, si els valors de l'estimador T_1 estan menys dispersos que els de T_2, respecte del valor θ.

Una imatge de les distribucions d'uns hipotètics estimadors pot ajudar-nos a il·lustrar aquests conceptes:

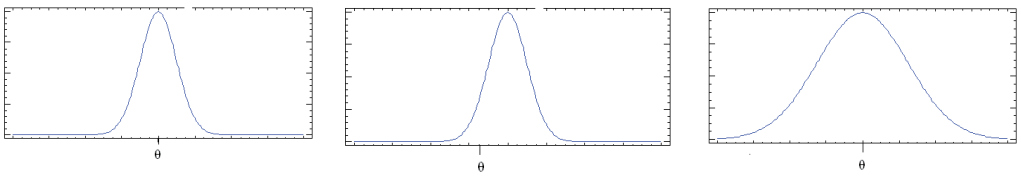


Figura a Figura b Figura c

Podem afirmar que els estimadors de les figures *a* i *c* no són esbiaixats de θ, mentre que el de la figura *b* sí que ho és. També es pot veure que l'estimador de *a* és més eficient que el de *c*.

Per a trobar estimadors que complisquen aquestes propietats, caldria explicar el mètode de màxima versemblança que ens ho asseguraria. Aquest estudi queda fora de l'objectiu del nostre treball.

Exemple 10

Provarem que l'estimador mitjana mostral $\overline{X}$ és un estimador no esbiaixat i consistent de μ. Per a demostrar-ho utilitzarem les propietats de l'esperança i la variància i el fet que cadascuna de les variables X_i es distribueixen amb el mateix model que la població amb mitjana μ i variància σ^2.

a) $\overline{X}$ és un estimador no esbiaixat de μ.

$$E[\overline{X}] = E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{\sum_{i=1}^{n} E[X_i]}{n} = \frac{n\mu}{n} = \mu.$$

b) $\overline{X}$ és un estimador consistent de μ, ja que:

$$\lim_{n\to\infty} E[\overline{X}] = \lim_{n\to\infty} \mu = \mu$$

$$\lim_{n\to\infty} \mathrm{Var}[\overline{X}] = \lim_{n\to\infty} Var\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} Var[X_i]}{n^2} = \lim_{n\to\infty} \frac{n\sigma^2}{n^2} = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0.$$

Exemple 11

Estudiem les propietats que compleixen els estimadors S^2_x, variància mostral i $S^2_{X_{n-1}}$, quasivariància mostral, com a estimadors del paràmetre σ^2.

a) S^2_x és un estimador esbiaixat i $S^2_{X_{n-1}}$ és un estimador no esbiaixat de σ^2.

$$S^2_X = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n} = \frac{\sum_{i=1}^{n} X_i^2}{n} - \overline{X}_n^2 \text{ amb } E[X_i] = \mu \quad \text{i} \quad Var[X_i] = \sigma^2$$

$$E[S^2_X] = E\left[\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_1^2 - \overline{X}^2\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_1^2] - E[\overline{X}^2] =$$

Per altra banda:

$$\begin{cases} \overbrace{Var[X_i]}^{\sigma^2} = E[X_i^2] - \overbrace{E[X_i]^2}^{\mu^2} \Rightarrow E[X_i^2] = \sigma^2 + \mu^2 \\ \underbrace{Var[\overline{X}]}_{\frac{\sigma^2}{n}} = E[\overline{X}^2] - \underbrace{E[\overline{X}]^2}_{\mu^2} \Rightarrow E[\overline{X}^2] = \frac{\sigma^2}{n} + \mu^2 \end{cases}$$

Substituïm aquests resultats en l'expressió anterior:

$$E[S_X^2] = \frac{1}{n}\sum_{i=1}^{n} E[X_1^2] - E[\overline{X}^2] = \frac{1}{n}\sum_{i=1}^{n}(\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \frac{\not{n}(\sigma^2 + \mu^2)}{\not{n}} - \left(\frac{\sigma^2}{n} + \mu^2\right) =$$

$$= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

Així, podem concloure que $E[S_X^2] = \frac{n-1}{n}\sigma^2$.

Estudiem ara l'esperança de l'altre estimador:

$$S_{X_{n-1}}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{n}{n-1}\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} = \frac{n}{n-1}S_{X_1}^2$$

$$E[S_{X_{n-1}}^2] = E\left[\frac{n}{n-1}S_X^2\right] = \frac{n}{n-1}E[S_X^2] = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2.$$

Hem demostrat, doncs, que S^2_x és un estimador esbiaixat i $S^2_{X_{n-1}}$ és un estimador no esbiaixat de σ^2.

b) Per a provar la consistència tots dos estimadors S^2_x i $S^2_{X_{n-1}}$, considerarem la distribució de l'estimador següent (teorema de Cochran), que ja vam veure en l'apartat 5 d'aquest mateix tema:

$\frac{nS_X^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2 \to \chi^2_{n-1}$ i com que coneixem l'esperança i la variància d'aquesta distribució, podem concloure que:

$$E\left[\frac{nS_X^2}{\sigma^2}\right] = n-1 \Leftrightarrow E[S_X^2] = \frac{n-1}{n}\sigma^2$$

$$Var\left[\frac{nS_X^2}{\sigma^2}\right] = 2(n-1) \Leftrightarrow \frac{n^2}{\sigma^4} Var\left[S_X^2\right] = 2(n-1) \Leftrightarrow Var\left[S_X^2\right] = \frac{2(n-1)}{n^2}\sigma^4$$

b.1) Podem demostrar fàcilment a partir dels resultats anteriors que la variància poblacional S_x^2 és un estimador consistent de σ^2.

$$\lim_{n\to\infty} E\left[\overline{X}\right] = \lim_{n\to\infty} \mu = \mu$$

$$\lim_{n\to\infty} Var\left[\overline{X}\right] = \lim_{n\to\infty} Var\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} Var\left[X_i\right]}{n^2} = \lim_{n\to\infty} \frac{n\sigma^2}{n^2} = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0.$$

b.2) Provem que la quasivariància $S^2_{X_{n-1}}$ és també un estimador consistent de σ^2. Per a això, comencem per demostrar el valor de la seua variància a partir de la de l'estimador S_x^2:

$$Var\left[S^2_{X_{n-1}}\right] = Var\left[\frac{n}{n-1}S_X^2\right] = \left(\frac{n}{n-1}\right)^2 Var\left[S_X^2\right] = \left(\frac{n}{n-1}\right)^2 \cdot \frac{2(n-1)}{n^2} \cdot \sigma^4 = \frac{2}{(n-1)}\sigma^4$$

$$\lim_{n\to\infty} E\left[S^2_{X_{n-1}}\right] = \lim_{n\to\infty} \sigma^2 = \sigma^2 \qquad \lim_{n\to\infty} Var\left[S^2_{X_{n-1}}\right] = \lim_{n\to\infty} \frac{2}{(n-1)}\sigma^4 = 0.$$

c) Comparem l'eficiència dels estimadors $S^2_{X_{n-1}}$ i S^2_x, mitjançant el quocient de les variàncies:

$$Var\left[S_X^2\right] = \frac{2(n-1)}{n^2}\sigma^4 \qquad Var\left[S^2_{X_{n-1}}\right] = \frac{2}{(n-1)}\sigma^4$$

$$\frac{Var\left[S_X^2\right]}{Var\left[S^2_{X_{n-1}}\right]} = \frac{\frac{2(n-1)}{n^2}\sigma^4}{\frac{2}{(n-1)}\sigma^4} = \frac{2(n-1)(n-1)}{2n^2} = \frac{(n-1)^2}{n^2} < 1 \Rightarrow Var\left[S_X^2\right] < Var\left[S^2_{X_{n-1}}\right].$$

Podem concloure que l'estimador S_x^2 és més eficient que $S^2_{X_{n-1}}$.

Com a conseqüència d'aquestes propietats, podem veure que les expressions algebraiques d'estimadors i estadístics, així com les fórmules dels intervals de confiança del tema següent, estan en funció de l'expressió de la quasivariància. Per a transformar un estimador en un altre tan sols cal recordar la relació que tenen entre si. $S^2_{X_{n-1}} = \frac{n}{n-1} S_X^2$, de fet és freqüent trobar bibliografia que utilitza l'una o l'altra a criteri de l'autor.

TEMA 9

Interferència: intervals de confiança

Objectius tema 9

- Estudiar i interpretar els intervals de confiança per a estimar els paràmetres poblacionals d'una i dues poblacions.
- Saber interpretar els resultats dels càlculs per a poder inferir conclusions.
- Saber calcular la grandària de la mostra necessària d'acord a les exigències de l'error que es pretenen en les estimacions.

9.1. Introducció

Si considerem una població la distribució de la qual és coneguda, però en desconeixem algun dels paràmetres, podem estimar el valor d'aquest paràmetre a partir d'una mostra representativa mitjançant un procés anomenat *estimació paramètrica*.

En la unitat anterior es va veure ja l'estimació puntual i en aquesta unitat estudiarem l'estimació per interval de confiança.

Per exemple, si estimem que el percentatge de vots d'un determinat partit polític és del 48% estarem realitzant una estimació puntual, mentre que si afirmem que l'estimació de vots es troba entre el 46% i el 59% amb un nivell de confiança del 95%, estarem considerant una estimació per interval de confiança.

Intervals de confiança

Quan ens plantegem estimar mitjançant intervals de confiança, pretenem trobar un interval en el qual puguem afirmar que el valor del paràmetre de la població que volem conèixer, i que podem denotar genèricament com a θ, hi està dins amb una alta probabilitat.

En primer lloc, cal fixar aquesta probabilitat, que anomenem *nivell de confiança* i que denotem per $1-\alpha$. El valor α s'anomena *nivell de significació*.

Si $I = (a, b)$ és l'interval que busquem, cal que $P(\theta \in I) = 1 - \alpha$, i es diu que I és l'interval de confiança per al paràmetre θ amb un nivell de confiança $1 - \alpha$.

Paràmetres poblacionals i mostrals

El paràmetre que cal estimar d'una població que hem anomenat θ serà, segons el cas, la mitjana μ, la variància σ^2, i en el cas de poblacions de Bernoulli, la proporció **p** que ens indica la probabilitat d'èxit, com ja vam veure en el tema de les distribucions.

Per a trobar aquest interval, començarem per triar una mostra representativa de la població d'una grandària *n*. Amb aquestes dades, en calcularem la mitjana $\overline{x}$, la quasivariància S^2_{n-1}, o la proporció **p** en el cas d'una població de Bernoulli.

Atès que els valors d'aquests paràmetres calculats a partir de les mostres (o una transformació adient per a cada cas, com hem vist en el tema anterior) segueixen una distribució coneguda, podem aplicar la teoria de la probabilitat, i amb l'ajuda de taules o programes estadístics, trobar els extrems de l'interval que necessitem en cada cas.

Error, grandària de la mostra i nivell de confiança

Per a començar a treballar amb un problema d'estimació cal fixar prèviament tres factors, que condicionaran el treball i la validesa del resultat:

- Nivell de confiança.
- Error.
- Grandària de la mostra.

Una vegada calculat l'interval de confiança podem mesurar l'error de l'estimació. Així, si $\theta \in I = (a, b)$, podem mesurar l'error amb l'expressió $E = \frac{b - a}{2}$. Com és obvi, aquest error depèn de l'amplitud de l'interval i també del nivell de confiança que demanem en l'estimació, ja que per al càlcul de l'interval es té en compte aquest nivell.

Si volem que aquest error disminuïsca, cal rebaixar el nivell de confiança. Recordem que el nivell de confiança $1 - \alpha$ és la probabilitat que el paràmetre θ que cal trobar estiga dins de l'interval de confiança, i en certa manera podem considerar-la com una mesura de garantia de la certesa de la nostra afirmació. En consqüència, és desitjable que el nivell de confiança siga un nombre proper a 1.

El tercer factor que juga en aquest equilibri és la grandària de la mostra n, ja que l'amplitud de l'interval –i, conseqüentment, l'error– disminueixen quan augmenta la grandària de la mostra. Així doncs, sempre podrem fixar, a priori, dos dels factors que cal controlar i obtenir el tercer en funció d'aquells.

Per altra part, en un experiment d'aquestes característiques en la vida real (control de qualitat d'un producte, previsions electorals, investigacions de biologia, d'ecologia, medicina, etc.), la manipulació dels elements de la mostra té un cost econòmic i de temps, per la qual cosa cal trobar un equilibri entre aquests tres factors que hem anomenat a fi de trobar una solució satisfactòria per la seua seguretat, l'exactitud i el cost per a tothom.

En els exemples que desenvoluparem al llarg del tema veurem aquesta relació aplicada en cada cas concret.

9.2. Intervals de confiança per als paràmetres d'una població

Per a utilitzar les fórmules següents en les quals intentarem estimar un interval de confiança per a un paràmetre poblacional, partirem de l'existència d'una mostra de grandària n de la qual coneixem, mitjançant els càlculs de l'estadística descriptiva que hem estudiat en els primers temes, la mitjana aritmètica $\overline{x}$ i la quasidesviació típica S_{n-1}. Aquests estadístics sempre són coneguts o podem calcular-los a partir dels elements de la mostra.

Respecte als paràmetres poblacionals que cal estimar, denotarem per μ la mitjana aritmètica i per σ la desviació típica.

En el cas d'una població de Bernoulli, denotarem per **p** la proporció obtinguda a partir dels valors de la mostra, i per ***p*** el paràmetre de la població que cal estimar.

En tots els casos denotarem per $1 - \alpha$ el nivell de confiança amb el qual decidirem treballar en cada cas. Els percentatges més habituals són del 90%, 95% i 99%.

Així mateix, denotarem per $z_{1-\frac{\alpha}{2}}$ el valor x de la variable normal tipificada, al qual correspon $P(Z \leq x) = 1 - \frac{\alpha}{2}$. Aquests valors, els podrem trobar en les taules que anirem especificant en cadascun dels casos, o també amb l'ajuda de qualsevol programa estadístic (Statgraphics, SPSS, R, etc.).

En el cas de les variables que en la notació necessiten que s'utilitze un subíndex per a expressar els graus de llibertat (khi quadrat, t de Student, F de Fisher-Snedecor), passarem a indicar aquesta probabilitat $1-\frac{\alpha}{2}$ o altres de semblants, amb un subíndex davant de la lletra que ens indica la variable. Així, denotarem per $t_{1-\frac{\alpha}{2}\,n-1}$ el valor x de la variable t de Student amb $n - 1$ graus de llibertat al qual correspon la probabilitat $P(t_{n-1} \leq x) = 1 - \frac{\alpha}{2}$, i raonarem de la mateixa manera en el cas de la variable khi quadrat amb $n - 1$ graus de llibertat χ^2_{n-1}, que denotarem per $\chi^2_{1-\frac{\alpha}{2}\,n-1}$.

Aquests aspectes es tractaran amb més detall en la resolució dels exemples de cadascun dels casos.

9.2.1. Intervals per a les mitjanes

En aquest apartat es definiran els intervals de confiança per a les mitjanes poblacionals i per a la diferència de mitjanes (si es tenen en compte dues poblacions) segons la informació de la qual es disposa sobre la població. En cada subapartat es presentarà l'interval de confiança i un exemple d'aplicació.

Interval per a la mitjana d'una població normal amb desviació típica poblacional coneguda σ

$$\mu \in \left(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

Demostrarem aquest interval com a exemple i la resta dels casos es poden resoldre seguint el mateix procés a partir de les distribucions dels estimadors que hem vist en el tema anterior.

Vam veure que la distribució de la mitjana mostral $\bar{x}$ és una distribució normal de mitjana μ i desviació típica $\frac{\sigma}{\sqrt{n}}$. Per tant, si tipifiquem, obtenim que la variable $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ es distribueix com una variable normal estandarditzada. Així, $Z \to N(0, 1)$. Per altra part, si s'observa la gràfica d'aquesta distribució, es pot afirmar que els valors centrals d'aquesta variable estan en un interval que té com a extrems uns percentils, que depenen del percentatge que volem considerar. Denotem per $1 - \alpha$ aquesta probabilitat.

Com podem veure en la gràfica següent, $z_{\frac{\alpha}{2}}$ i $z_{1-\frac{\alpha}{2}}$ seran aquests percentils, del quals podem dir que $P\left(Z \leq z_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$ i que $P\left(Z \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$.

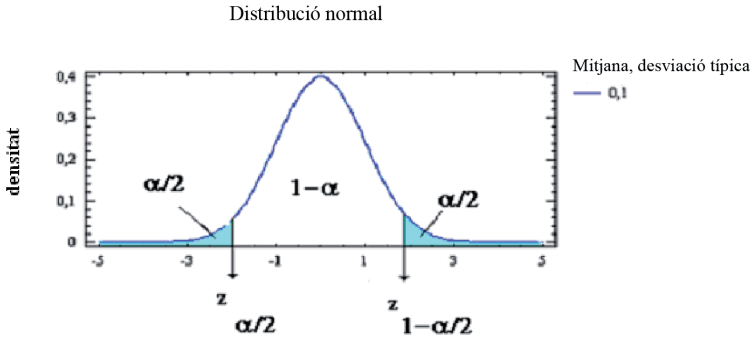


Figura 1

Per la simetria de la distribució, es compleix que $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$, per la qual cosa, es pot afirmar que $P\left(-z_{1-\frac{\alpha}{2}} \leq Z \leq z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$ i substituint l'expressió de la variable Z en aquest interval:

$$P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}}\right) = 1-\alpha.$$

En realitat, si $\bar{x}$ és una mitjana mostral, com que és un valor de la variable $\overline{X}$, sabem que, almenys en un $(1 - \alpha)$% dels casos, es compleix que:

$$-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}}.$$

Aïllem pas a pas el paràmetre μ en aquesta desigualtat:

$$-z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{x}-\mu \leq z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$-\bar{x}-z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{x}+z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\bar{x}-z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}+z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}.$$

Així, l'interval de confiança de la mitjana amb un nivell de confiança de l'$(1 - \alpha)$% és:

$$\mu \in \left(\bar{x}-z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x}+z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right).$$

Exemple 1

Volem fer un estudi del nombre d'hores que dormen els estudiants d'una determinada universitat.

Definim X = «nombre d'hores de son diàries» com la variable que cal estudiar i sabem que aquesta es distribueix seguint un model normal de mitjana desconeguda, però per altres treballs podem considerar que la seua desviació típica és coneguda i el seu valor és 3.

Per a estimar aquest valor de μ caldrà triar una mostra. Per a això, farem una enquesta a 25 alumnes triats aleatòriament i després de preguntar-los el nombre d'hores de son de cadascun el dia de l'enquesta, obtenim una mitjana de 7 hores. Aquesta dada, la denotarem per $\bar{x} = 7$ i cal distingir que aquest és un paràmetre mostral, ja que l'hem calculat amb les dades obtingudes en l'enquesta dels individus de la mostra.

Hem decidit fer una estimació mitjançant intervals, treballant amb un nivell de confiança del 95%.

Així doncs, necessitem aplicar la fórmula anterior, ja que cal estimar el paràmetre mitjana poblacional, una vegada coneguda la desviació típica ($\sigma = 3$):

$$\mu \in \left(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right).$$

Cal insistir en el fet que aquesta desviació és un paràmetre poblacional que ja coneixem per treballs anteriors i no obtingut a partir de les dades de l'enquesta de la mostra. Així: podem observar que coneixem tots els valors que cal substituir-hi llevat del valor $z_{1-\frac{\alpha}{2}}$. Per calcular-lo començarem per deduir fàcilment el valor d'$(1-\frac{\alpha}{2})$, ja que com hem dit abans, $1 - \alpha$ és el nivell de confiança que establim a priori per a fer el treball:

$$1-\alpha = 0{,}95 \quad \rightarrow \quad \alpha = 0{,}05 \quad \rightarrow \quad \frac{\alpha}{2} = 0{,}025 \quad \rightarrow \quad 1-\frac{\alpha}{2} = 0{,}975.$$

Tanmateix, busquem el valor de la variable *x* de tal manera que $P(Z \leq x) = 0{,}975$. És, en realitat, el valor d'un percentil, que podem obtenir de la taula de la funció de distribució de la variable normal tipificada. Aquest valor és $x = 1{,}96$, per la qual cosa, utilitzant la notació d'aquest tema, $z_{1-\frac{\alpha}{2}} = 1{,}96$ o també, $z_{0{,}975} = 1{,}96$.

I ara substituirem tots els valors en la fórmula per a obtenir l'interval:

$$\mu \in \left(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) = \left(7 - 1{,}96\frac{3}{\sqrt{25}}, 7 + 1{,}96\frac{3}{\sqrt{25}} \right) = (5{,}824,\ 8{,}176).$$

Podem concloure que, amb una probabilitat del 95%, la mitjana del nombre d'hores de son diàries dels alumnes de la universitat està dins d'aquest interval.

Atès que l'amplitud de l'interval és 8,176 – 5,824 = 2,352 hores, l'error en la nostra estimació és de $E = \frac{b-a}{2} = \frac{2{,}352}{2} = 1{,}176$ hores.

Si aquest error, el considerem excessiu, podem comprovar la diferència del que obtindríem si la nostra mostra augmentara a 30 individus. Suposem que el valor de la mitjana mostral haguera estat el mateix $\bar{x} = 7$:

$$\mu \in \left(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) = \left(7 - 1{,}96 \frac{3}{\sqrt{30}}, 7 + 1{,}96 \frac{3}{\sqrt{30}} \right) = (5{,}926,\ 8{,}074).$$

També podem observar l'efecte que es produeix en l'amplitud de l'interval si rebaixem el nivell de confiança al 90%, per exemple, i mantenim la grandària de la mostra en $n = 25$ per a contrastar l'efecte d'aquest darrer canvi.

En aquest cas caldrà tornar a buscar en les taules el valor de $z_{1-\frac{\alpha}{2}}$:

$$1 - \alpha = 0{,}9 \quad \rightarrow \quad \alpha = 0{,}1 \quad \rightarrow \quad \frac{\alpha}{2} = 0{,}05 \quad \rightarrow \quad 1 - \frac{\alpha}{2} = 0{,}95.$$

Així, busquem en la taula de la distribució acumulada de la variable normal tipificada aquesta probabilitat i obtenim que $z_{1-\frac{\alpha}{2}} = 1{,}65$, la qual cosa ens dóna l'interval:

$$\mu \in \left(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) = \left(7 - 1{,}65 \frac{3}{\sqrt{25}}, 7 + 1{,}65 \frac{3}{\sqrt{25}} \right) = (6{,}01,\ 7{,}99).$$

Com es pot observar, aquest interval és més menut i la resposta és més ajustada, però les conclusions han perdut nivell de confiança.

Nota

Si la grandària de la mostra n és suficientment gran, podem utilitzar la fórmula d'aquest apartat, considerant que la desviació típica poblacional σ, encara que fóra desconeguda, coincideix amb el valor de la quasidesviació típica de la mostra S_{n-1}. Per tractar-se d'un manual de docència, estimarem adient aquesta consideració per a mostres de grandària n superiors a 30.

Interval per a la mitjana d'una població normal amb desviació típica poblacional desconeguda

En aquest cas l'interval de confiança per a la mitjana és:

$$\mu \in \left(\bar{x} - {}_{1-\frac{\alpha}{2}} t_{n-1} \frac{S_{n-1}}{\sqrt{n}}, \bar{x} + {}_{1-\frac{\alpha}{2}} t_{n-1} \frac{S_{n-1}}{\sqrt{n}} \right).$$

Exemple 2

Volem confirmar el funcionament d'una màquina envasadora de fruita en almívar, per comprovar que el pes dels pots envasats d'1 kg s'ajusta a l'etiquetatge i a la normativa. Per a dur a terme el treball vam seleccionar 20 pots triats aleatòriament i vam calcular-ne la mitjana aritmètica dels pesos, que és $\bar{x}$ = 995 g i la seua desviació típica, S = 5 g.

Volem estimar el pes mitjà dels pots envasats, per la qual cosa utilitzarem l'interval per a la mitjana μ poblacional, però cal observar que no coneixem el valor de la desviació típica de la població.

També cal considerar que la grandària de la mostra, n = 20, és inferior a 30 i no podem estimar aquesta desviació σ basant-nos en el valor de la quasidesviació típica mostral S_{n-1}. En aquest cas, doncs, emprarem la fórmula següent:

$$\mu \in \left(\bar{x} - {}_{1-\frac{\alpha}{2}}t_{n-1}\frac{S_{n-1}}{\sqrt{n}}\,,\ \bar{x} + {}_{1-\frac{\alpha}{2}}t_{n-1}\frac{S_{n-1}}{\sqrt{n}} \right).$$

En les estimacions treballarem amb el 99% de nivell de confiança.

Com podem veure, cal calcular la quasidesviació típica S_{n-1}, a partir del valor que coneixem de S = 5 i n = 20.

$$S_{n-1} = \sqrt{\frac{n}{n-1}}S = \sqrt{\frac{20}{19}}5 = 5{,}130.$$

També cal esbrinar el valor d'${}_{1-\frac{\alpha}{2}}t_{n-1}$ a la taula de la variable t de Student o amb l'ajuda d'algun programa. En primer lloc, calcularem el valor d'$(1-\frac{\alpha}{2})$, ja que:

$$1-\alpha = 0{,}99 \ \rightarrow\ \alpha = 0{,}01 \ \rightarrow\ \frac{\alpha}{2} = 0{,}005 \ \rightarrow\ 1-\frac{\alpha}{2} = 0{,}995,$$

i en la taula de la distribució acumulada de la variable t de Student amb 19 graus de llibertat, trobem el valor de la variable x de tal manera que $P(t_{19} \leq x) = 0{,}995$. Aquest valor és x = 2,861, per la qual cosa, utilitzant la notació d'aquest tema ${}_{1-\frac{\alpha}{2}}t_{n-1} = {}_{0{,}995}t_{19} = 2{,}861$.

I substituint aquests valors en la fórmula, obtenim que:

$$\mu \in \left(\bar{x} - {}_{1-\frac{\alpha}{2}}t_{n-1}\frac{S_{n-1}}{\sqrt{n}}\,,\ \bar{x} + {}_{1-\frac{\alpha}{2}}t_{n-1}\frac{S_{n-1}}{\sqrt{n}} \right) = \left(995 - 2{,}861\frac{5{,}130}{\sqrt{20}}\,,\ 995 + 2{,}861\frac{5{,}130}{\sqrt{20}} \right) =$$
$$(991{,}718,\ 998{,}282).$$

Aquest interval ens indica que la màquina no està ben ajustada, ja que el valor d'1 kg = 1.000 g no està dins de l'interval. La conclusió que podríem extraure'n és que la mitjana del pes dels pots envasats té un 99% de probabilitat de ser lleugerament inferior a 1 kg.

Podem comentar que si la mostra fóra de 100 pots i els valors de la mitjana i desviació típica coincidiren amb els anteriors, podríem haver aplicat la fórmula de l'apartat anterior. Ja hem indicat que si la mostra és gran, la quasidesviació típica mostral $S_{n\text{-}1}$, podrem considerar-la una bona estimació de la desviació típica poblacional σ.

Així podem veure que els resultats canviarien un poc. Apliquem-ho:

$$\mu \in \left(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) = \left(995 - 2{,}576 \ \frac{5{,}130}{\sqrt{100}}, 995 + 2{,}576 \ \frac{5{,}130}{\sqrt{100}} \right) =$$
$$= (993{,}679,\ 996{,}321).$$

Podem veure que els resultats varien una mica però hem arribat a la mateixa conclusió.

Us convidem a comprovar que en el primer cas i amb una mostra de 20 pots, arribaríem a la mateixa conclusió si treballem amb el 95% de nivell de confiança, ja que al disminuir aquest, l'amplitud de l'interval disminueix i els valors de la mitjana queden encara més allunyats del valor de 1.000 g.

En aquest cas, atès que $1 - \alpha = 0{,}95$, cal consultar en les taules aquest valor de la variable *t* de Student, ${}_{1-\frac{\alpha}{2}}t_{n-1} = {}_{0{,}975}t_{19} = 2{,}093$ i amb la resta de les dades conegudes s'obté l'interval:

$$\mu \in \left(\bar{x} - {}_{1-\frac{\alpha}{2}}t_{n-1} \frac{S_{n-1}}{\sqrt{n}}, \bar{x} + {}_{1-\frac{\alpha}{2}}t_{n-1} \frac{S_{n-1}}{\sqrt{n}} \right) = \left(995 - 2{,}093 \frac{5{,}130}{\sqrt{20}}, 995 + 2{,}093 \frac{5{,}130}{\sqrt{20}} \right) =$$
$$= (992{,}599,\ 997{,}401),$$ per la qual cosa la nostra conclusió no canvia.

9.2.2. Altres intervals

Interval per a la proporció d'una població de Bernoulli

L'interval de confiança per a la proporció, en aquest cas, és:

$$p \in \left(p - z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}, p + z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \right).$$

Exemple 3

Per a estudiar la proporció de persones interessades a adquirir un cert producte nou que es va llançar al mercat, es fa una enquesta a 100 persones, 87 de les quals han respost afirmativament. Calculem un interval de confiança per a la quota de mercat de la població amb un nivell de significació de l'1%, si suposem que els individus que han donat resposta a l'enquesta són una mostra representativa de la població.

Entenem que la quota de mercat que cal estimar és la proporció d'individus de la població p que estaran interessats a adquirir el nou producte, per la qual cosa utilitzem la fórmula anterior.

Per a obtenir l'interval plantejat fem servir una mostra de grandària $n = 100$ i a partir de les respostes, podem calcular la proporció en la mostra $p = \frac{87}{100} = 0{,}87$.

Com que el plantejament és treballar amb un nivell de significació de l'1%, això implica que $\alpha = 0{,}01 \rightarrow \frac{\alpha}{2} = 0{,}005 \rightarrow 1 - \frac{\alpha}{2} = 0{,}995$, per la qual cosa caldrà obtenir de la taula de la distribució acumulada de la normal tipificada, el valor $z_{1-\frac{\alpha}{2}} = 2{,}58$, que substituirem en la fórmula anterior:

$$p \in \left(p - z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}, p + z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \right) =$$

$$= \left(0{,}87 - 2{,}58 \sqrt{\frac{0{,}87 \cdot 0{,}13}{100}}, 0{,}87 + 2{,}58 \sqrt{\frac{0{,}87 \cdot 0{,}13}{100}} \right) = (0{,}783, 0{,}958),$$

és a dir, amb un nivell de confiança del 99% podem estimar que la proporció de la població que està interessada a adquirir el nou producte està entre el 78,3% i el 95,8%.

Interval per a la variància d'una població normal

En aquest cas l'interval de confiança per a la variància és:

$$\sigma^2 \in \left(\frac{(n-1)S^2_{n-1}}{{}_{1-\frac{\alpha}{2}}\chi^2_{n-1}}, \frac{(n-1)S^2_{n-1}}{{}_{\frac{\alpha}{2}}\chi^2_{n-1}} \right).$$

Exemple 4

Aprofundirem en l'exemple 2 dels apartats anteriors, i farem una estimació de la variància dels pesos de la màquina envasadora de pots de fruita.

Per a això, considerarem els mateixos valors mostrals: la grandària de la mostra $n = 20$, la mitjana aritmètica dels pesos $\bar{x} = 995$ g i la desviació típica $S = 5$ g.

Estimarem un interval per a la variància σ^2 amb un nivell de confiança del 95%, per a la qual cosa utilitzarem la fórmula d'abans.

Necessitem calcular la quasivariància de la mostra i la podem obtenir a partir de la variància:

$$S^2_{n-1} = \frac{n}{n-1} S^2 = \frac{20}{19} 25 = 26{,}316.$$

En aquest cas el nivell de confiança escollit és del 95%, és a dir, $1 - \alpha = 0{,}95$:

$$1-\alpha = 0{,}95 \quad \rightarrow \quad \alpha = 0{,}05 \quad \rightarrow \quad \frac{\alpha}{2} = 0{,}025 \quad \rightarrow \quad 1-\frac{\alpha}{2} = 0{,}975.$$

i en la taula de la distribució acumulada de la variable khi quadrat amb 19 graus de llibertat, trobem els valors de ${}_{1-\frac{\alpha}{2}}\chi^2_{n-1} = {}_{0{,}975}\chi^2_{19}$ i de ${}_{\frac{\alpha}{2}}\chi^2_{n-1} = {}_{0{,}025}\chi^2_{19}$, ja que com que la variable no és simètrica, obtenim dos valors diferents en valor absolut per a cada probabilitat.

En el nostre cas obtenim ${}_{0{,}975}\chi^2_{19} = 32{,}8523$ i per a ${}_{0{,}025}\chi^2_{19} = 8{,}90652$. Si ara substituïm tots aquests valors en la fórmula de l'interval obtenim:

$$\sigma^2 \in \left(\frac{(n-1)S^2_{n-1}}{{}_{1-\frac{\alpha}{2}}\chi^2_{n-1}}, \frac{(n-1)S^2_{n-1}}{{}_{\frac{\alpha}{2}}\chi^2_{n-1}} \right) = \left(\frac{19 \cdot 26{,}316}{32{,}8523}, \frac{19 \cdot 26{,}316}{8{,}90652} \right) = (15{,}220,\ 56{,}139),$$

Per a comparar aquest paràmetre amb el valor mostral de l'enunciat, podem calcular l'interval per a estimar la desviació típica dels pesos dels pots, tan sols calculant l'arrel quadrada dels extrems de l'interval anterior:

$$\sigma \in \left(\sqrt{15.220}\ , \sqrt{56.139}\right) = (3{,}901,\ 7{,}492).$$

Així, podem estimar amb un nivell de confiança del 95% que la desviació típica dels pesos envasats en la màquina té un valor que està dins del darrer interval que hem trobat.

9.3. Intervals de confiança per als paràmetres de dues poblacions

Fins ara hem abordat estimacions dels paràmetres d'una població i hem pogut calcular intervals dins dels quals es troba el valor del paràmetre que cal estimar (mitjana, proporció, variància) amb una probabilitat que anomenem *nivell de confiança* i que, generalment, pren valors per damunt del 90%.

En aquest apartat, compararem els paràmetres de dues poblacions. Per a dur a terme el treball, necessitem les dades de dues mostres diferents, extretes de les seues respectives poblacions, ja que volem inferir conclusions sobre els valors dels seus paràmetres poblacionals.

Cal detallar que en aquests casos no podem estimar els valors dels paràmetres de les poblacions, més aviat podem comparar-los i concloure quin dels dos és més alt i en quina mesura. És a dir, podem estimar l'interval en el qual es mouen els valors de la seua diferència, i en els cas concret de la comparació de variàncies, podem estimar l'interval corresponent als valors dels seus quocients. Aquesta circumstància ens permetrà estimar quin és més alt i fins a quin punt és significativa la diferència.

Pel que fa a la notació, hem utilitzat el subíndex X per als valors mostrals i poblacionals d'una de les poblacions que cal considerar i el subíndex Y per a l'altra.

Cal observar que en alguns casos les dades de totes dues poblacions estaran referides a la mateixa variable que cal estudiar i voldrem extraure conclusions de la magnitud de les diferències en el valor dels paràmetres que cal comparar. Per exemple, si volem estimar si les qualificacions mitjanes de l'assignatura d'estadística del grup del matí i del grup de la vesprada són significativament diferents,

podrem denotar per $\bar{x}$ la nota mitjana dels alumnes del matí que hem seleccionat aleatòriament per a la mostra i per S_x^2, la variància d'aquestes dades. Denotarem per $\bar{y}$ i per S_y^2 els mateixos paràmetres referits a les qualificacions dels estudiants triats del grup de la vesprada. És clar que les dades que cal treballar corresponen també a la variable «qualificacions en l'assignatura d'estadística».

En altres casos, voldrem comparar les diferències en els paràmetres de dues variables amb les dades obtinguedes dels individus d'una mateixa població. Per exemple, podríem estudiar si hi ha més dispersió en les dades de l'alçada o en les del pes dels nounats d'un hospital. En aquest cas podríem parlar de la comparació de variàncies de la variable *X*, que seria el pes, i la variable *Y*, que seria l'alçada de cada xiquet.

Si seguim amb la qüestió de la notació, mantenim el paral·lelisme amb els valors de les mitjanes μ_x i μ_y per a cadascuna de les poblacions, i els símbols σ_x^2 i σ_Y^2 per a denotar les variàncies.

En el cas d'una població de Bernoulli, denotem per p_x i p_y les proporcions obtingudes a partir dels valors de les mostres, i per $\mathbf{p_x}$ i $\mathbf{p_y}$ les proporcions poblacionals que cal estimar de les respectives poblacions.

Quant a les notacions dels valors de les variables, totes coincideixen amb les de l'apartat anterior, tret de la variable F de Fisher-Snedecor, amb n –1 i m – 1 graus de llibertat, que denotarem per $f_{(n-1),(m-1)}$. Cal recordar que l'ordre d'aquests subíndexs no és commutatiu.

La resta d'aspectes són iguals als tractats en l'apartat anterior, i anirem comentant-los amb els exemples de cada cas.

És important ressaltar, en aquest apartat, que ara estudiem, una certa particularitat a l'hora d'interpretar els resultats de l'interval trobat (ens hi referirem amb més profunditat en cadascun dels casos). Així mateix, seria adient fixar-se en el cas del quocient de variàncies per la diferència amb la resta dels intervals en aquest aspecte.

9.3.1. Intervals per a la diferència de mitjanes

Per a començar, abordarem el cas de la diferència de mitjanes i caldrà saber si estem en el cas de mostres independents de dues poblacions o si es tracta de mostres relacionades o aparellades.

En el primer cas, no cal que la grandària d'ambdues mostres coincidisca. No obstant això, és obvi que en el cas de les mostres aparellades les grandàries d'ambdues han de ser iguals, ja que a cada individu, li corresponen un parell d'observacions, les quals permeten definir una variable que cal treballar, anomenada *diferència D*. Aquesta nova variable s'obté com la diferència de les dues observacions de cada element, així: $d_i = x_i - y_i$.

Interval per a la diferència de mitjanes de poblacions independents amb variàncies poblacionals conegudes

L'interval de confiança per a la diferència de mitjanes en aquest cas és:

$$\mu_x - \mu_y \in \left(\bar{x} - \bar{y} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_Y^2}{m}}, \bar{x} - \bar{y} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_Y^2}{m}} \right).$$

Exemple 5

Volem comparar l'eficiència de dues empreses de missatgeria internacional atenent al temps que tardem a rebre les remeses. Sabem que el nombre d'hores que tarden a arribar els enviaments de l'empresa A segueix una distribució normal de la qual coneixem la desviació típica de 25 hores, i el de l'empresa B també segueixen una distribució normal amb desviació típica de 30 hores.

Per a estudiar la situació, anotem en 10 enviaments de l'empresa A, un temps mitjà de 80 hores, mentre que en la mostra dels 15 enviaments de l'empresa B el temps mitjà és de 75 hores.

Estimarem, amb un interval al 99% de nivell de confiança, quina empresa té una mitjana inferior utilitzant la fórmula que hem presentat anteriorment:

$$\mu_x - \mu_y \in \left(\bar{x} - \bar{y} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_Y^2}{m}}, \bar{x} - \bar{y} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_Y^2}{m}} \right).$$

Per a la població A, podem anotar el paràmetre de la població $\sigma_x^2 = 25^2 = 625$, i amb les dades de la mostra $n = 10$, $\bar{x} = 80$.

Per a la població B, podem anotar el paràmetre de la població $\sigma_Y^2 = 30^2 = 900$, i amb les dades de la mostra $m = 15$, $\bar{y} = 75$.

Com que treballem amb un nivell de confiança del 99%, podem calcular $1 - \alpha = 0{,}99$ i calculem el valor d'$1-\frac{\alpha}{2}$ ja que:

$$1-\alpha = 0{,}99 \quad \rightarrow \quad \alpha = 0{,}01 \quad \rightarrow \quad \frac{\alpha}{2} = 0{,}005 \quad \rightarrow \quad 1-\frac{\alpha}{2} = 0{,}995.$$

i en la taula de la distribució acumulada de la variable normal tipificada, busquem el valor de la variable x de tal manera que $P(Z \leq x) = 0{,}995$. Aquest valor és $x = 2{,}57583$ per la qual cosa, utilitzant la notació d'aquest tema, $z_{1-\frac{\alpha}{2}} = z_{0{,}995} = 2{,}57583$.

Substituïm aquestes dades en la fórmula i obtenim que:

$$\mu_x - \mu_y \in \left(\bar{x} - \bar{y} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_Y^2}{m}}, \bar{x} - \bar{y} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_Y^2}{m}} \right) =$$

$$= \left(80 - 75 - 2{,}576 \sqrt{\frac{625}{10} + \frac{900}{15}},\ 80 - 75 + 2{,}576 \sqrt{\frac{625}{10} + \frac{900}{15}} \right) = (-23{,}511,\ 33{,}511).$$

En tots els intervals en què ens plantegem la diferència de paràmetres cal considerar tres possibilitats per a interpretar la solució. Suposem que l'interval solució és: $\mu_x - \mu_y \in (a, b)$.

- Si $0 \in (a, b)$, interpretarem que la diferència entre els paràmetres no és significativa i podrem estimar que $\mu_x \approx \mu_y$.
- Si $a > 0$, i $b > 0$, interpretarem que la diferència és positiva amb una alta probabilitat, per la qual cosa podrem estimar que $\mu_x - \mu_y > 0 \rightarrow \mu_x > \mu_y$.
- Si $a < 0$, i $b < 0$, interpretarem que la diferència és negativa amb una alta probabilitat, per la qual cosa podrem estimar que $\mu_x - \mu_y < 0 \rightarrow \mu_x > \mu_y$.

En els dos darrers casos els valors dels intervals ens donaran una idea aproximada dels valors al voltant dels quals oscil·la la diferència entre els paràmetres.

En el nostre exemple, com que el valor 0 està dins de l'interval solució, estimarem que la diferència de 5 hores que hi ha entre les mitjanes mostrals no és significativa per a estimar que el temps mitjà dels enviaments és inferior en una altra empresa, i considerarem que totes dues són igual d'eficaces.

Interval per a la diferència de mitjanes de poblacions independents amb variàncies poblacionals desconegudes però iguals

L'interval de confiança per a la diferència de mitjanes en aquest cas és:

$$\mu_x - \mu_y \in \left(\bar{x} - \bar{y} - {}_{1-\frac{\alpha}{2}} t_{n+m-2} \sqrt{\frac{(n-1) S^2_{X_{n-1}} + (m-1) S^2_{Y_{n-1}}}{n+m-2}} \sqrt{\frac{1}{n} + \frac{1}{m}}, \right.$$

$$\left. \bar{x} - \bar{y} + {}_{1-\frac{\alpha}{2}} t_{n+m-2} \sqrt{\frac{(n-1) S^2_{X_{n-1}} + (m-1) S^2_{Y_{n-1}}}{n+m-2}} \sqrt{\frac{1}{n} + \frac{1}{m}} \right)$$

Aplicarem aquesta fórmula quan no coneguem les variàncies poblacionals σ^2_x ni σ^2_y, però puguem afirmar que són iguals. Si no és el cas, podrem fer un interval per a estimar la igualtat entre totes dues variàncies, com es veurà un poc més endavant en aquest mateix apartat, i si la interpretació ens permet estimar que són iguals continuarem calculant aquesta diferència de mitjanes.

Exemple 6

En una mesura de control de qualitat en la fabricació d'unes peces, volem comparar si dos processos de producció són equivalents i si mantenen els mateixos estàndards de qualitat. Considerarem que les variàncies de totes dues poblacions són iguals.

Per a realitzar el treball, agafem unes quantes peces de cada línia i les classifiquem amb l'ajuda d'un índex de qualitat que resumeix la informació de diversos indicadors. Les dades de les mostres figuren en la taula següent:

Línia X	10	9	7	6	12	3	7	9	10	6		
Línia Y	12	8	5	11	9	10	13	7	12	9	8	13

Amb aquesta informació volem estimar, mitjançant un interval de confiança, si la qualitat mitjana és la mateixa en tots dos processos de fabricació. Treballarem amb un nivell de confiança del 95%.

Per a dur a terme els càlculs, identificarem els valors dels paràmetres de cadascuna de les mostres, ja que com diu l'enunciat, suposarem que $\sigma_x^2 = \sigma_Y^2$.

De la mostra de la població *X*, anotarem $n = 10$, $\bar{x} = 7{,}9$ i $S^2_{x_{n-1}} = 6{,}77$.

De la mostra de la població *Y*, anotarem $m = 12$, $\bar{y} = 9{,}75$ i $S^2_{Y_{n-1}} = 6{,}39$.

En les taules podem trobar la dada que ens falta $t_{1-\frac{\alpha}{2}\,n+m-2}$ i, com que treballem amb un nivell de confiança del 95%, tenim que $1 - \alpha = 0{,}95$, i a partir d'això obtenim que:

$$1-\alpha = 0{,}95 \quad \rightarrow \quad \alpha = 0{,}05 \quad \rightarrow \quad \frac{\alpha}{2} = 0{,}025 \quad \rightarrow \quad 1-\frac{\alpha}{2} = 0{,}975.$$

Així, cal calcular en les taules de la distribució acumulada de la variable *t* de Student amb 20 graus de llibertat, el valor de la variable $t_{1-\frac{\alpha}{2}\,n+m-2} = t_{0{,}975\,20} = 2{,}086$ i ara substituïm aquests valors en la fórmula:

$$\mu_x - \mu_y \in \left(\bar{x} - \bar{y} - t_{1-\frac{\alpha}{2}\,n+m-2} \sqrt{\frac{(n-1)S^2_{x_{n-1}} + (m-1)S^2_{Y_{m-1}}}{n+m-2}} \sqrt{\frac{1}{n}+\frac{1}{m}}, \right.$$

$$\left. \bar{x} - \bar{y} + t_{1-\frac{\alpha}{2}\,n+m-2} \sqrt{\frac{(n-1)S^2_{x_{n-1}} + (m-1)S^2_{Y_{m-1}}}{n+m-2}} \sqrt{\frac{1}{n}+\frac{1}{m}} \right) =$$

$$\left(7{,}9 - 9{,}75 - 2{,}086 \sqrt{\frac{9 \cdot 6{,}77 + 11 \cdot 6{,}39}{20}} \sqrt{\frac{1}{10}+\frac{1}{12}}, \right.$$

$$\left. 7{,}9 - 9{,}75 + 2{,}086 \sqrt{\frac{9 \cdot 6{,}77 + 11 \cdot 6{,}39}{20}} \sqrt{\frac{1}{10}+\frac{1}{12}} \right) =$$

$= (-4{,}138\ ,\ 0{,}438)$.

Podem concloure que com que el 0 està dins d'aquest interval, la diferència entre les mitjanes no és significativa i que totes dues línies de producció tenen un nivell de qualitat semblant, cosa que podem afirmar amb un nivell de confiança del 95%.

Interval per a la diferència de mitjanes de dues poblacions amb mostres relacionades

En aquest cas cal definir prèviament la variable $D = X - Y$, per a la qual denotarem per $S^2_{D_{n-1}}$ la quasivariància mostral, on $d_i = x_i - y_i$. La grandària de totes dues mostres coincideix necessàriament i la denotarem per n. L'interval de confiança per a la diferència de mitjanes en aquest cas és:

$$\mu_x - \mu_y \in \left(\bar{x} - \bar{y} - {}_{1-\frac{\alpha}{2}} t_{n-1} \sqrt{\frac{S^2_{D_{n-1}}}{n}}, \bar{x} - \bar{y} + {}_{1-\frac{\alpha}{2}} t_{n-1} \sqrt{\frac{S^2_{D_{n-1}}}{n}} \right).$$

Exemple 7

Per a millorar el grau de satisfacció dels clients del banc A s'ha plantejat eliminar la major part de les comissions que aquest cobrava als seus clients per alguns serveis. Per a avaluar l'eficàcia del dit plantejament, s'ha passat a 8 clients una enquesta dissenyada per a esbrinar el grau mitjà de satisfacció abans i després de l'eliminació de les comissions en una escala de 0 a 3. Els resultats es detallen en la taula que presentem a continuació:

Abans	1,2	1,3	1,5	1,4	1,7	1,9	1,4	1,2
Després	1,4	1,7	1,5	1,3	2	2,1	1,7	1,6

Calculem un interval de confiança al 95% per veure si la mesura adoptada pel banc ha donat resultat.

Per a això necessitem calcular els valors de la variable $D = X - Y$, restant els valors emparellats tal com es veu en la taula següent:

X = valors abans de la mesura	1,2	1,3	1,5	1,4	1,7	1,9	1,4	1,2
Y = valors després de la mesura	1,4	1,7	1,5	1,3	2	2,1	1,7	1,6
$D = X - Y$	-0,2	-0,4	0	0,1	-0,3	-0,2	-0,3	-0,4

Amb la calculadora, esbrinem els valors següents de les mostres de les variables X, Y, com ara $n = 8$, $\bar{x} = 1{,}45$, $\bar{y} = 1{,}6625$, i amb les dades de la variable D, calculem $S^2_{D_{n-1}} = 0{,}0327$.

Com que treballem amb un nivell de confiança del 95%:

$$1-\alpha = 0{,}95 \quad \rightarrow \quad \alpha = 0{,}05 \quad \rightarrow \quad \frac{\alpha}{2} = 0{,}025 \quad \rightarrow \quad 1-\frac{\alpha}{2} = 0{,}975.$$

Així, en les taules de la distribució acumulada de la variable t de Student cal calcular ${}_{1-\frac{\alpha}{2}}t_{n-1} = {}_{0{,}975}t_7 = 2{,}3646$. Amb la resta de les dades conegudes substituïm aquests valors i s'obté l'interval:

$$\mu_x - \mu_y \in \left(\bar{x} - \bar{y} - {}_{1-\frac{\alpha}{2}}t_{n-1}\sqrt{\frac{S^2_{D_{n-1}}}{n}}, \bar{x} - \bar{y} + {}_{1-\frac{\alpha}{2}}t_{n-1}\sqrt{\frac{S^2_{D_{n-1}}}{n}} \right) =$$

$$= \left(1{,}45 - 1{,}6625 - 2{,}3646\sqrt{\frac{0{,}0327}{8}},\ 1{,}45 - 1{,}6625 + 2{,}3646\sqrt{\frac{0{,}0327}{8}} \right) = (-0{,}3637, -0{,}613).$$

Com que els dos valors dels extrems de l'interval són negatius, podem inferir que $\mu_x - \mu_y < 0 \rightarrow \mu_x < \mu_y$, per la qual cosa interpretem que la mitjana de les enquestes de satisfacció és més gran en les que han estat realitzades després de la mesura establerta per satisfer els clients. Podem estimar que la mesura d'eliminar comissions sí que ha aconseguit l'objectiu per al qual havia estat dissenyada.

9.3.2. Altres intervals de confiança

Interval per a la diferència de proporcions de dues poblacions de Bernoulli

L'interval de confiança per a la diferència de proporcions en aquest cas és:

$$p_x - p_y \in \left(p_x - p_y - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_x(1-p_x)}{n} + \frac{p_Y(1-p_Y)}{m}},\ p_x - p_y + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_x(1-p_x)}{n} + \frac{p_Y(1-p_Y)}{m}} \right).$$

Exemple 8

En un control de qualitat en la maquinària d'una empresa, volem conèixer si les dues màquines que tenim són igual d'eficaces, i considerem la proporció de peces defectuoses que ixen de cadascuna en el procés d'elaboració. Per a fer el treball seleccionem aleatòriament una mostra de 200 peces de la màquina A, 15 de les quals són defectuoses, i 250 peces de la màquina B, 16 de les quals també ho són.

Si volem treballar amb un nivell de confiança del 95%, calcularem l'interval anterior amb les dades del nostre problema:

En primer lloc calculem les dades que necessitem de la mostra de la màquina A:

$$n = 200 \qquad p_x = \frac{15}{200} = 0{,}075.$$

i les dades de la mostra de la màquina B:

$$m = 250 \qquad p_Y = \frac{16}{250} = 0{,}064.$$

Com que treballarem amb un nivell de confiança del 95%, podem calcular el valor d'$1-\frac{\alpha}{2}$, ja que $1-\alpha = 0{,}95 \;\rightarrow\; \alpha = 0{,}05 \;\rightarrow\; \frac{\alpha}{2} = 0{,}025 \;\rightarrow\; 1-\frac{\alpha}{2} = 0{,}975$ i en la taula de la distribució acumulada de la variable normal tipificada, busquem el valor de la variable x que faça complir que $P(Z \le x) = 0{,}975$. Aquest valor és $x = 1{,}96$ que denotem per $z_{1-\frac{\alpha}{2}} = 1{,}96$.

Substituïm aquestes dades en la fórmula de l'interval i obtenim que:

$$p_x - p_y \in \left(p_x - p_y - z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_x(1-p_x)}{n} + \frac{p_Y(1-p_Y)}{m}},\; p_x - p_y + z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_x(1-p_x)}{n} + \frac{p_Y(1-p_Y)}{m}} \right) =$$

$$\left(0{,}075 - 0{,}064 - 1{,}96 \sqrt{\frac{0{,}075 \cdot 0{,}925}{200} + \frac{0{,}064 \cdot 0{,}936}{250}},\; 0{,}075 - 0{,}064 + 1{,}96 \sqrt{\frac{0{,}075 \cdot 0{,}925}{200} + \frac{0{,}064 \cdot 0{,}936}{250}} \right) =$$

$= (-0{,}036,\ 0{,}058)$.

La conclusió que podem extraure d'aquest interval és que treballant amb un nivell de confiança del 95%, la diferència entre les proporcions mostrals que hem obtingut no és significativa; podem estimar que la proporció de peces defectuoses en la producció de totes dues màquines és la mateixa. Aquesta afirmació, la basem en el fet que el valor 0 està dins de l'interval, ja que, per tractar-se d'una diferència, té les mateixes conclusions que els casos que hem treballat anteriorment sobre la diferència de mitjanes.

Interval per al quocient de dues variàncies de dues poblacions normals

En aquest cas l'interval de confiança per al quocient de variàncies és:

$$\frac{\sigma_X^2}{\sigma_Y^2} \in \left(\frac{\frac{(n-1)S_{X_{n-1}}^2}{n}}{\frac{(m-1)S_{Y_{n-1}}^2}{m}} \cdot \frac{1}{f_{(n-1),(m-1)}^{1-\frac{\alpha}{2}}},\; \frac{\frac{(n-1)S_{X_{n-1}}^2}{n}}{\frac{(m-1)S_{Y_{n-1}}^2}{m}} \cdot \frac{1}{f_{(n-1),(m-1)}^{\frac{\alpha}{2}}} \right).$$

Exemple 9

Per a estudiar aquest interval, prendrem com a mostra les dades de l'exemple 6, on per a comparar les mitjanes de dues poblacions s'utilitzen les dues mostres que presentem a continuació.

Recordem que en aquell apartat ja vam comentar que es tractava de dues mostres on necessitàvem pressuposar que ambdues poblacions tenien la mateixa variància. Si no podem afirmar aquesta condició per treballs anteriors, caldrà començar per fent el treball que presentem a continuació. En cas que la inferència ens permeta estimar que són iguals, podrem portar a terme el treball que ja vam fer en l'exemple 6.

Recordem que es tractava d'unes quantes peces agafades de dues línies de producció i que s'havien classificat amb l'ajuda d'un índex de qualitat que resumeix la informació de diversos indicadors. Les dades de les mostres figuren en la taula següent:

Línia X	10	9	7	6	12	3	7	9	10	6		
Línia Y	12	8	5	11	9	10	13	7	12	9	8	13

Amb aquesta informació volem estimar, mitjançant un interval de confiança, si la variabilitat en la qualitat de totes dues línies de producció és la mateixa. Entenem per variabilitat el valor de les variàncies, que és el paràmetre utilitzat com a indicador de la dispersió, ja que volem saber si els indicadors de qualitat permeten comprovar que estan igual de propers a la seua mitjana en tots dos processos.

Per dur a terme els càlculs, identificarem els valors dels paràmetres de cadascuna de les mostres:

De la mostra de la població X, anotarem $n = 10$, $\overline{x} = 7{,}9$, $S^2_{X_{n-1}} = 6{,}77$.
De la mostra de la població Y, anotarem $m = 12$, $\overline{y} = 9{,}75$, $S^2_{Y_{n-1}} = 6{,}39$.

Treballarem amb un nivell de confiança del 95%, per la qual cosa $1 - \alpha = 0{,}95$ i així:

$$1-\alpha = 0{,}95 \quad \rightarrow \quad \alpha = 0{,}05 \quad \rightarrow \quad \frac{\alpha}{2} = 0{,}025 \quad \rightarrow \quad 1-\frac{\alpha}{2} = 0{,}975.$$

Buscarem en la taula de la distribució acumulada de la variable F de Fisher-Snedecor, els percentils que corresponen als valors de $\frac{\alpha}{2} = 0{,}025$ i $1-\frac{\alpha}{2} = 0{,}975$. Cal comentar que si utilitzem taules cal trobar el primer valor en funció del segon que podrem trobar en les taules que solen publicar-se, aplicant la propietat de la funció de distribució d'aquesta variable F de Fisher-Snedecor que expliquem.

Així, en les taules o amb un programa estadístic trobem el valor del percentil $_{1-\frac{\alpha}{2}}f_{(n-1),(m-1)}=_{0,975}f_{9,1}$ = 3,5879 i per a calcular l'altre valor fem ús de la propietat $_{\frac{\alpha}{2}}f_{(n-1),(m-1)}=\frac{1}{_{1-\frac{\alpha}{2}}f_{(m-1),(n-1)}}$, que aplicada al nostre cas seria $_{0,025}f_{9,1}=\frac{1}{_{0,975}f_{1,9}}=$ $=\frac{1}{3,91207}=0,255619.$

Si substituïm tots aquests valors en l'expressió de l'interval obtindrem com a resultat que:

$$\frac{\sigma_X}{\sigma_X}\in\left(\frac{\frac{(n-1)S^2_{X_{n-1}}}{n}}{\frac{(m-1)S^2_{Y_{n-1}}}{m}}\cdot\frac{1}{_{1-\frac{\alpha}{2}}f_{(n-1),(m-1)}},\frac{\frac{(n-1)S^2_{X_{n-1}}}{n}}{\frac{(m-1)S^2_{Y_{n-1}}}{m}}\cdot\frac{1}{_{\frac{\alpha}{2}}f_{(n-1),(m-1)}}\right)=$$

$$\left(\frac{\frac{9}{10}6,77}{\frac{11}{12}6,39}\cdot\frac{1}{3,5879},\frac{\frac{9}{10}6,77}{\frac{11}{12}6,39}\cdot\frac{1}{0,255619}\right)=(0,2899,\ 4,0694).$$

Aquests valors dels extrems de l'interval ens permeten inferir que les variàncies de les poblacions de les quals provenen les mostres són iguals, és a dir, els diferents productes de cadascuna de les línies de fabricació que comparem, presenten el mateix comportament en la dispersió dels valors de qualitat respecte a la mitjana de cadascuna d'aquelles.

Aquesta inferència està basada en el fet que el valor 1 està dins de l'interval calculat, ja que com que estem comparant les variàncies mitjançant la seua ràtio o quocient, els valors dels extrems ens diuen que els quocients estan al voltant de la unitat, i aquesta solució permet inferir igualtat entre les dues mesures comparades per divisió.

En general, quan comparem les variàncies de dues poblacions mitjançant l'expressió de l'interval que hem mostrat en aquest apartat, podem arribar a tres possibles resultats, la interpretació dels quals comentem a continuació, sempre tenint en compte que es tracta de l'anàlisi d'uns quocients.

Si considerem que l'interval solució és $\frac{\sigma^2_X}{\sigma^2_Y}\in(a,\ b)$.

- Si $1\in(a,\ b)$, interpretarem que la diferència entre els paràmetres no és significativa i podrem estimar que $\sigma^2_X\approx\sigma^2_Y$.
- Si $a>1$ i $b>1$, interpretarem que el quocient és superior a 1 amb una alta probabilitat, per la qual cosa podrem estimar que $\frac{\sigma^2_X}{\sigma^2_Y}>1\ \rightarrow\ \sigma^2_X>\sigma^2_Y$.

- Si $a < 1$ i $b < 1$, interpretarem que el quocient és inferior a 1 en una alta probabilitat, per la qual cosa podrem estimar que $\frac{\sigma^2_X}{\sigma^2_Y} < 1 \quad \rightarrow \quad \sigma^2_X > \sigma^2_Y$.

En els dos darrers casos els valors dels intervals ens donaran una idea aproximada dels valors al voltant dels quals oscil·la el quocient entre els paràmetres.

9.4. Error i grandària de la mostra

Ja havíem comentat en la introducció que quan es fa un treball d'inferència, hi ha tres factors que determinen la possible solució. Recordem-los:

- Nivell de confiança.
- Error de la solució.
- Grandària de la mostra.

En els exemples que hem treballat al llarg del desenvolupament del tema, hem vist que el nivell de confiança expressa la probabilitat que l'estimació siga correcta, determina els valors que hem trobat en les taules de les funcions de distribució de les diferents variables treballades i és costum utilitzar-hi valors que van del 90% al 99%, i com més gran l'agafem, més gran és l'amplitud de l'interval.

Seria desitjable que l'interval solució fóra el més ajustat possible als valors que volem inferir. L'expressió que tenim de l'error és aquesta desviació del valor central. Així, direm que l'error és $E = \frac{b-a}{2}$ i, per tant, està directament relacionat amb l'amplitud de l'interval.

El tercer factor que cal considerar és la grandària de la mostra. Com més gran és aquesta, més exactes són els resultats. Així doncs, podem plantejar-nos la qüestió a l'inrevés, per a fer un treball que satisfaça les nostres necessitats.

És possible plantejar-nos una situació com la següent: considerem el treball de l'exemple 1, on ja vam veure l'efecte que comporten un canvi en el nivell de confiança i un canvi en la grandària de la mostra. Estudiem ara l'error en la solució del plantejament inicial que teníem amb una mostra de 25 alumnes, treballant amb un nivell de confiança del 95%. El resultat que hi vam obtenir, considerant $\sigma = 3$, va ser:

$$\mu \in \left(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) = \left(7 - 1{,}96 \frac{3}{\sqrt{25}}, 7 + 1{,}96 \frac{3}{\sqrt{25}} \right) = (5{,}824,\ 8{,}176).$$

Podem concloure que, amb una probabilitat del 95%, la mitjana del nombre d'hores de son diàries dels alumnes de la universitat està dins d'aquest interval.

Atès que l'amplitud de l'interval és 8,176 – 5,824 = 2,532 hores, l'error en l'estimació és $E = \frac{b-a}{2} = \frac{2{,}352}{2} = 1{,}176$.

Si volem acotar aquest error i que siga de 0,5 hores, podem esbrinar, a piori, la grandària de la mostra amb el que hem de treballar i així conèixer el nombre d'individus al qual cal fer l'enquesta. Així:

$$E = \frac{b-a}{2} = \frac{\left(\bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) - \left(\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)}{2} = \frac{2z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}}{2} = z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} = 1{,}96\frac{3}{\sqrt{n}} = 0{,}5.$$

D'on podem aïllar el valor de n, $\sqrt{n} = \frac{1{,}96 \cdot 3}{0{,}5} = 11{,}76 \rightarrow n = 11{,}76^2 = 138{,}30$, és a dir, amb 139 individus podem extraure l'interval amb l'error que ens havíem plantejat.

9.5. Problemes proposats

En aquest epígraf es plantejaran un conjunt de problemes per a la resolució dels quals és necessari conèixer la teoria desenvolupada al llarg de la unitat.

Exercici 1

Per a millorar els beneficis mitjans de l'empresa A s'ha posat en marxa un procés de renovació. Per a avaluar-lo, anotarem els beneficis mitjans de totes les seccions de què consta l'empresa, abans i després de la modernització, que es detallen a continuació:

Abans	1,2	1,3	1,5	1,4	1,7	1,9	1,4	1,2
Després	1,4	1,7	1,5	1,3	2	2,1	1,7	1,6

Calcula un interval de confiança al 95% per veure si ha estat justificat o no el procés de renovació. Explica la interpretació del resultat i raona la resposta.

Exercici 2

La regidoria de joventut d'un ajuntament treballa amb la dada que l'edat a què s'independitzen els joves és una variable normal amb mitjana 29 anys i desviació típica 3 anys. Encara que la desviació típica no planteja dubtes, sí que se sospita que la mitjana ha descendit, a causa de la política d'ajuda a l'ocupació que ha dut a terme l'ajuntament. Per justificar la sospita s'ha fet un estudi amb 100 joves que s'acaben d'independitzar i s'hi ha obtingut una mitjana de 28,1 anys. Estima al 99% de nivell de confiança si la sospita és certa.

Exercici 3

Suposem que el nombre de pulsacions dels homes de 20 a 25 anys pot considerar-se que segueix una distribució normal de mitjana 72 pulsacions/min i desviació típica de 9 pulsacions/min. Si una mostra de 100 corredors de marató dóna una mitjana de 64 pulsacions, podem estimar amb un nivell de confiança del 99% que en aquests esportistes la disminució de la mitjana respecte a la població és significativa?

Exercici 4

En una factoria de fabricació de taulells es mantenen dues línies de producció diferents. Volem comparar si ambdues línies tenen els mateixos nivells de qualitat, per a aquest fi es disposa de les dades d'un índex de qualitat que engloba diferents indicadors. Hem obtingut aquestes dades en cadascuna de les línies A i B.

A	10	9	7	6	12	3	7	9	10	6
B	12	8	5	11	9	10	13	7	8	13

Amb la informació de què disposem, podem afirmar, amb un nivell de confiança del 95%, que la qualitat mitjana és la mateixa en ambdues línies de producció? Justifica la resposta utilitzant un interval de confiança i considerant que $\sigma_X^2 = \sigma_Y^2$.

Exercici 5

Per a introduir un material en línia fem un estudi previ i enquestem 100 alumnes. Si 25 ens donen una opinió favorable del producte, calcula amb un interval de confiança al 95% el percentatge d'alumnes que podem esperar que utilitzaran el material.

Exercici 6

Volem comparar la mitjana de temps que necessiten dues empreses de missatgeria per a complir els encàrrecs i així contractar aquella que ens done millor servei. Per a això, anotem el temps emprat en 14 serveis de l'empresa A i obtenim una mitjana de 17 hores amb una desviació típica d'1,22 hores. També analitzem 25 serveis de la companyia B i obtenim una mitjana de 19 hores amb una desviació típica d'1,34 hores. Suposem que els temps per a totes dues empreses es distribueixen normalment.

a) Calcula un interval de confiança al 90% per a comparar les variàncies.
b) Podem afirmar que l'empresa B té una mitjana de temps millor en els seus serveis?

TEMA 10

Contrastos d'hipòtesi

Objectius tema 10

- Conèixer les fases i assimilar la nomenclatura dels contrastos.
- Saber interpretar els resultats dels contrastos per a poder inferir conclusions.
- Conèixer els contrastos d'alguns paràmetres d'una població.
- Conèixer els contrastos per a comparar alguns paràmetres de dues poblacions.

10.1. Introducció

Es pot definir el contrast d'hipòtesi com la part de la inferència estadística que té com a objectiu comprovar, mitjançant mètodes matemàtics, hipòtesis realitzades sobre el valor d'algun paràmetre d'una o diverses poblacions.

Comencem plantejant-nos algunes situacions:

- Un fabricant de piles afirma que la durada mitjana de les seues piles és de 53 hores com a mínim i la desviació típica, de 4 hores. Rebutjaríem aquesta afirmació en el cas que una pila durara 48 hores? I si la mitjana de la durada de 100 piles fóra de 50 hores? I si aquesta mitjana fóra de 56 hores?
- Tenim un dau que suposem correcte. El llancen 100 vegades i obtenim 25 vegades el número 5. Podrem dir que el dau és correcte? O cal verificar que no està trucat en funció dels resultats?
- Suposem dues empreses que produeixen les piles anteriors. Al llarg dels anys aquestes piles han tingut una durada similar, però en l'actualitat la segona empresa afirma que les seues piles duren més perquè ha fet una millora en la producció. Podem creure aquesta afirmació?

En la pràctica és freqüent trobar-nos davant situacions com aquestes en les quals cal prendre decisions sobre hipòtesis estadístiques senzilles relatives als paràmetres d'una població o de dues. És per això que aquesta branca de la inferència també s'anomena *teoria de la decisió*.

És clar que, en prendre una decisió estadística, mai estarem completament segurs d'encertar. Al llarg del tema veurem com prendre aquesta decisió i com minimitzar el risc d'equivocar-nos.

10.2. Contrastos d'hipòtesi

Un test estadístic és un procediment per a extraure, a partir d'una mostra aleatòria i representativa, conclusions que permeten acceptar o rebutjar una hipòtesi prèviament emesa sobre el valor d'un paràmetre desconegut d'una població.

Si volem conèixer la veracitat d'una informació, que considerarem que és una hipòtesi sobre la població, la contrastarem amb la informació que traurem d'una mostra. Si ambdues informacions coincideixen dins d'un marge que considerarem admissible, mantindrem que la hipòtesi inicial era certa. En cas contrari, la rebutjarem i seria lògic plantejar-nos noves hipòtesis que explicaren les dades observades.

Bàsicament, podríem comparar aquests mètodes amb un judici en què, en principi, se suposa la innocència de l'acusat (hipòtesi nul·la) i es tracta d'aportar proves per a rebutjar aquesta hipòtesi. La innocència serà rebatuda quan les proves ho demostren amb un alt grau de fiabilitat.

10.2.1. Hipòtesi nul·la i hipòtesi alternativa

Començarem plantejant-nos les dues primeres situacions que hem presentat en començar el tema. En ambdós exemples hi ha una hipòtesi de partida i uns resultats obtinguts a partir de les dades d'una mostra, que difereixen de la hipòtesi. Ens preguntem si aquesta diferència és atribuïble a l'atzar o si estem davant una hipòtesi de partida que era falsa.

En el primer exemple, evidentment, la primera qüestió estaria fora de lloc. No ens semblarà raonable que el fet que un element contradiga una generalització, ens porte a pensar que estem davant d'una informació falsa. Sempre refutarem la hipòtesi inicial sobre la població i obtindrem dades reals d'una mostra representativa de la població. Analitzarem la segona qüestió, on tenim una mostra de 100 piles que tenen una durada mitjana de 50 hores.

	Cas 1: piles	**Cas 2: dau**
Hipòtesi	El fabricant afirma que $\mu = 53, \sigma = 4$	Si el dau és correcte, la proporció de cincs és $p = 1/6 = 0{,}167$
Resultats de la mostra	$\overline{X} = 50$	**p** = 0,25
Interrogant	La diferència observada pot atribuir-se a l'atzar? Podem suposar raonable que la mostra ha estat extreta de la població sobre la qual hem fet la hipòtesi?	

En ambdós exemples hem considerat una hipòtesi que, en principi, admetem com a vàlida i que desitgem contrastar. S'anomena *hipòtesi nul·la* i es designa per H_o i cal definir la hipòtesi contrària, que s'anomena *hipòtesi alternativa* i es designa per H_{11}, que serà admesa quan el resultat del contrast siga rebutjat H_o.

	Cas 1: piles	**Cas 2: dau**
Hipòtesi	$\begin{cases} H_0: \mu = 53 \\ H_1: \mu \neq 53 \end{cases}$	$\begin{cases} H_0: p = 0{,}167 \\ H_1: p \neq 0{,}167 \end{cases}$

En aquest tema abordarem contrastos paramètrics, on la hipòtesi nul·la tracta sobre el valor d'un paràmetre de la població. Aquests contrastos poden tenir dos tipus d'hipòtesis:

- Hipòtesis simples: especifiquen un únic valor del paràmetre. Exemple: $\theta = \theta_0$.
- Hipòtesis compostes: especifiquen un interval de valors. Exemple: $\theta < \theta_0$.

Nosaltres realitzarem contrastos en els quals la hipòtesi nul·la serà simple i l'alternativa serà composta.

En principi, podem plantejar tres tipus d'hipòtesis alternatives:

- Prova bilateral $\begin{cases} H_o: = \theta_o \\ H_1: \neq \theta_o \end{cases}$ que plantejarem en general si no tenim informació de quina de les dues possibilitats portarà que H_o siga falsa.

- Prova unilateral $\begin{cases} H_o: = \theta_o \\ H_1: > \theta_o \end{cases}$ per la dreta, i $\begin{cases} H_o: = \theta_o \\ H_1: < \theta_o \end{cases}$ per l'esquerra, que, plantejarem quan sapiem en quina possibilitat està la probabilitat més alta que la hipòtesi nul·la siga falsa o quan no ens interessa si és falsa en l'altra direcció. (Si volem comprovar que un tractament millora un cert indicador de malaltia, donem per fet que és evident que ja coneixem que no l'empitjora.)

10.2.2. Tipus d'error

Al prendre qualsevol decisió, com a conseqüència de l'aplicació d'un test estadístic, pot ocórrer que:

1. Acceptem H_o i aquesta siga vertadera.
2. Acceptem H_o i aquesta siga falsa.
3. Rebutgem H_o i aquesta siga vertadera.
4. Rebutgem H_o i aquesta siga falsa.

H_o	Acceptem	Rebutgem
vertadera	No hi ha error	**Error de tipus I**
falsa	**Error de tipus II**	No hi ha error

En els casos 1 i 4 estem prenent una decisió encertada. En els casos 2 i 3 estem prenent una decisió equivocada, és a dir, fem un error.

Quan rebutgem H_o i aquesta és veritat, direm que fem un error de tipus I.
Quan acceptem H_o i aquesta és falsa, direm que fem un error de tipus II.

La probabilitat de fer un error de tipus I es denota per α i es denomina *nivell de significació del contrast d'hipòtesi*. Aquest nivell de significació és una probabilitat menuda (0,1; 0,05; 0,01), que habitualment tria l'investigador.

La probabilitat de fer un error de tipus II, es denota per β. El nombre $1 - \beta$ es diu *potència del contrast d'hipòtesi*. Aquesta disminueix quan augmenta la grandària de la mostra.

Un bon test, cal dissenyar-lo minimitzant les probabilitats de fer-hi errors. Però no és fàcil reduir tots dos tipus d'errors simultàniament. En general, quan disminueix un tipus d'error, augmenta l'altre, llevat que augmentem la grandària de la mostra.

Per a obtenir un bon contrast d'hipòtesi plantejarem l'elecció adient de la hipòtesi nul·la, de tal manera que l'error de tipus I siga el de «pitjors conseqüències». Per exemple, si estem estudiant si un nou fàrmac produeix efectes secundaris o no en els pacients, seria correcte plantejar-ho així:

H_o: el fàrmac produeix efectes secundaris,
H_1: el fàrmac no produeix efectes secundaris,

ja que és més greu concloure que el fàrmac no produeix efectes secundaris si els produeix (error de tipus I), que decidir que els produeix i que no siga veritat (error de tipus II).

10.3. Disseny d'un contrast d'hipòtesi

Desenvoluparem el procés complet per a realitzar un contrast d'hipòtesi. Detallarem cadascuna de les fases i utilitzarem l'exemple 1 per a anar concretant cada concepte.

1. Identificació de les hipòtesis i les dades

Aquest apartat ja està suficientment desenvolupat en l'apartat anterior.

En el cas de la duració de les piles volem comprovar que la mitjana és de 53 hores. També coneixem que $\sigma = 4$. Per a contrastar aquesta afirmació triem una mostra de 100 piles i n'obtenim el valor de la mitjana mostral $\overline{X} = 50$.

Així plantejarem les hipòtesis: $\begin{cases} H_o: \mu_o = 53 \\ H_1: \mu_o \neq 53 \end{cases}$. Podem, doncs, concloure que farem un contrast bilateral sobre el valor de la mitjana.

2. Tria d'un estadístic de contrast T

Aquest és un estimador del paràmetre poblacional, del qual volem refutar el valor, i del qual coneixem la distribució. Suposem que es tracta d'estimadors no esbiaixats respecte del paràmetre del qual es planteja la hipòtesi nul·la.

En cadascun dels casos que anem plantejant en els apartats següents, presentarem l'estadístic adient i la distribució, que ja vam explicar en els apartats corresponents del tema 8. Abordarem un exemple de cada cas, com en els temes anteriors.

En l'exemple 1 de les piles, utilitzarem com a estadístic de contrast, la variable $Z = \dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$, de la qual coneixem que es distribueix seguint el model d'una normal tipificada.

3. Determinació de la regió crítica o de significació i de la regió d'acceptació

Si la hipòtesi nul·la és certa, l'estadístic de contrast és una variable de la qual coneixem la distribució. Ens plantegem esbrinar quins seran els valors més probables d'aquest i quins seran molt improbables.

Busquem un interval, dins del qual serà molt probable que queden inclosos la majoria dels valors possibles de l'estadístic, i que anomenarem *regió d'acceptació*. S'hi dóna aquest nom perquè si el valor de l'estadístic mostral hi està dins, podem confirmar que la hipòtesi nul·la és certa. La informació extreta de la mostra no contradiu l'afirmació inicial que donàvem per vàlida respecte al valor del paràmetre poblacional.

Al contrari, si aquest valor de l'estadístic calculat amb les dades de la mostra cau fora de l'interval de la regió d'acceptació, és a dir, en la regió crítica, rebutjarem la hipòtesi nul·la. Podem plantejar-nos una disjuntiva:

1. El valor trobat és molt improbable, però gràcies a l'atzar la nostra mostra aleatòria ha donat un valor molt allunyat dels valors més probables.
2. El valor extret de la realitat està molt lluny dels valors més probables de la distribució, que per ser coneguda podem calcular, per la qual cosa podem concloure que la distribució que havíem donat com a vàlida no és correcta.

Per a això, definirem el nivell de significació α, el qual és la probabilitat que triem per a definir l'amplitud de la regió d'acceptació, ja que considerarem que la probabilitat que un estadístic tinga un valor dins d'aquest interval és $1 - \alpha$. Aquesta elecció de α també ens defineix la regió crítica R, ja que considerarem que $P(T \in R / \theta = \theta_0)$. És evident que ha de ser menuda. Valors habituals són 0,10, 0,05 i 0,01.

Podem comentar que el nivell de significació és la probabilitat de fer un error de tipus I, és a dir, de rebutjar la hipòtesi nul·la quan aquesta és certa.

En l'exemple de les piles, la hipòtesi nul·la estava plantejada sota el supòsit que la mitjana poblacional de la duració de les piles era de 53 hores i la desviació típica, de 4 hores. Ara bé, la mitjana de la mostra de 100 piles és de 50 hores.

Amb aquestes dades calcularem les regions de crítica i d'acceptació que assignarem al nostre contrast amb l'expressió de l'estadístic $Z = \dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$, que es distribueix com una variable normal tipificada. Treballarem amb un nivell de significació $\alpha = 0{,}05$.

Calcularem els extrems de l'interval (a, b) de la regió d'acceptació, considerant que la variable Z és simètrica i que hem plantejat una prova bilateral.

$$P(a \leq Z \leq b) = 0{,}95 \quad \rightarrow \quad P(Z \leq b) = 0{,}975.$$

Per a trobar $b = Z_{0{,}975} = 1{,}96$ utilitzarem taules o programes estadístics.

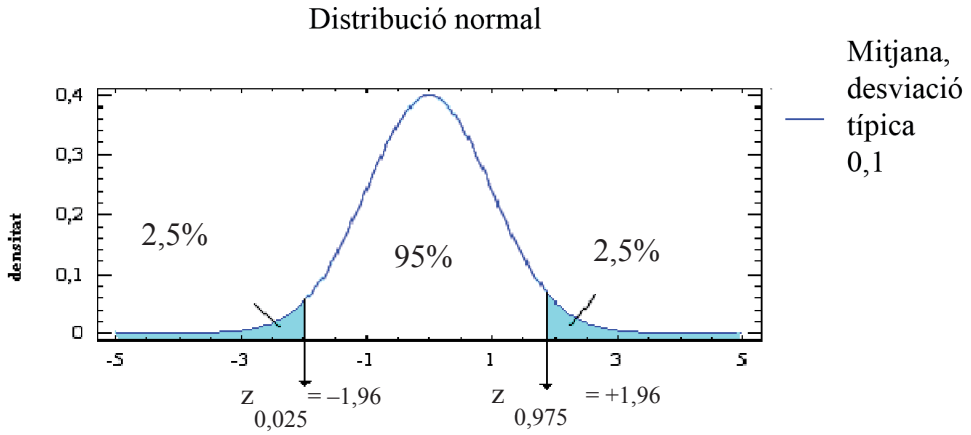


Figura 1. *Prova bilateral*

Així, la regió d'acceptació serà l'interval (–1,96; +1,96) i la resta de valors reals seran la regió crítica. Es pot observar en el gràfic de la figura 1.

En general, quan es tracta d'una prova bilateral, la regió crítica està repartida en dos intervals, a l'esquerra i la dreta de l'interval que defineix la zona central determinada per la regió d'acceptació.

En la prova unilateral, la regió d'acceptació queda en un lateral i la regió crítica és un únic interval que queda a la seua dreta o esquerra, segons el cas. És per aquesta raó que es diu *prova unilateral* o *d'una cua*, enfront de la bilateral o de dues cues.

4. Càlcul del valor de l'estadístic i elaboració de les conclusions

Amb les dades dels paràmetres que donem per veritables i amb els quals hem definit la hipòtesi nul·la i amb els valors mostrals extrets de la mostra representativa, triada adequadament, calculem el valor de l'estadístic en cada cas.

En el nostre exemple, l'estadístic val $Z = \dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}} = \dfrac{50 - 53}{\dfrac{4}{\sqrt{100}}} = \dfrac{-3}{0{,}4} = -7{,}5$.

Si el valor de l'estadístic cau dins de la regió d'acceptació, acceptem la hipòtesi nul·la i la considerem vertadera. Si, al contrari, cau dins de la regió crítica, rebutjarem la hipòtesi nul·la, i la suposarem falsa.

En el nostre exemple, és evident que el valor de l'estadístic està fora de la regió d'acceptació i, en aquest cas, molt allunyat. Conseqüentment, rebutjarem la hipòtesi nul·la que afirmava que la duració mitjana de les piles és de 53 hores.

10.4. Contrastos d'hipòtesi per als paràmetres d'una població

Per a construir aquests tests cal utilitzar les distribucions dels estadístics que vam veure en el tema 8. Seguirem la mateixa notació i necessitarem les mateixes condicions que les estudiades en tots els apartats d'estimació.

Considerarem $X_1, X_2, \ldots X_n$ una mostra aleatòria d'una població que es distribueix amb qualsevol model de distribució amb mitjana μ i variància σ^2.

Explicarem el procés de cada cas amb un exemple. En cada apartat proposarem tots tres tipus de contrast: prova bilateral, prova unilateral per la dreta i prova unilateral per l'esquerra. Abordarem les tres possibilitats en els diferents exemples, de manera que hi quedaran prou clares les semblances i les diferències entre si.

Com que la inferència mitjançant l'estimació per intervals i el contrast d'hipòtesi estan molts relacionats, en cada cas abordarem els mateixos exemples del tema anterior, però aplicant-hi un plantejament adient a la teoria de la decisió.

10.4.1. Contrast d'hipòtesi per a mitjanes

Contrast d'hipòtesi per a la mitjana mostral (coneguda σ^2 poblacional)

Estadístic $T = \dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ que es distribueix $T \to N(0, 1)$

1. Prova bilateral $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$ $R = \left\{T: |T| \geq z_{1-\frac{\alpha}{2}}\right\}$

2. Prova unilateral per la dreta $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$ $R = \left\{T: T \geq z_{1-\alpha}\right\}$

3. Prova unilateral per l'esquerra $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$ $R = \left\{T: T \leq -z_{1-\alpha}\right\}$

Exemple 1

Recordem el plantejament de l'estudi de les hores de son diàries dels estudiants universitaris. En sabem, per estudis anteriors, que la mitjana del nombre d'hores de son és de 7,5 hores amb una desviació típica de 3 hores. Per a corroborar aquesta afirmació, triem una mostra de 25 estudiants i en calculem la mitjana mostral $\overline{x} = 7$.

Plantejarem una prova unilateral per l'esquerra, ja que les sospites són que la mitjana puga haver disminuït. Treballarem amb un nivell de significació del 5%.

$$\begin{cases} H_0: \mu = 7{,}5 \\ H_1: \mu < 7{,}5 \end{cases} \qquad R = \{T: T \leq -z_{1-\alpha}\}.$$

Calculem l'estadístic amb les dades mostrals i poblacionals de l'enunciat:

$$T = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{7 - 7{,}5}{\frac{3}{\sqrt{25}}} = \frac{-0{,}5}{0{,}6} = -0{,}8333.$$

Per a trobar la regió crítica, calcularem el valor d'$(1 - \alpha)$ i el percentil corresponent de la variable Z normal tipificada:

$$\alpha = 0{,}05 \quad \rightarrow \quad 1 - \alpha = 0{,}95.$$

Per a trobar el valor de $z_{1-\alpha}$, calcularem el valor de x que fa complir que $P(Z \leq x) = 0{,}95$, amb ajuda de les taules o d'un programa informàtic. Així, $z_{1-\alpha} = 1{,}64$.

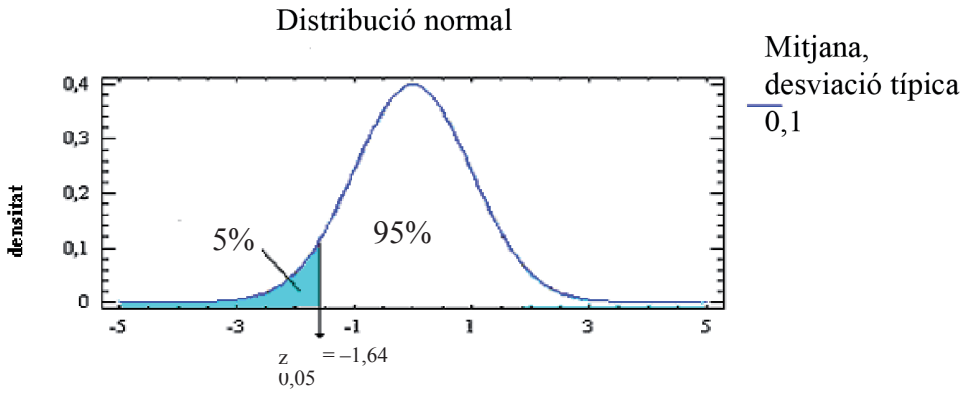


Figura 2. *Prova unilateral per l'esquerra*

$$R = \{T: T \leq -z_{1-\alpha}\} = \{T: T \leq -1{,}64\} = (-\infty, -1{,}64)$$

Podem comprovar que $Z = -0{,}8333 \notin R$, per tant no rebutjarem la hipòtesi nul·la i podem donar per bona l'afirmació que el nombre d'hores de son dels universitaris té una mitjana de 7,5 hores. Gràficament es pot observar a la figura 2.

Contrast d'hipòtesi per a la mitjana mostral (si no es coneix σ^2 poblacional)

Estadístic $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$, podrem afirmar que $T \to t_{n-1}$ és a dir, la variable T té una distribució t de Student amb $n - 1$ graus de llibertat. Recordem que S fa referència a la quasivariància mostral.

1. Prova bilateral $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$ $R = \left\{T : |T| \geq {}_{1-\frac{\alpha}{2}}t_{n-1}\right\}$

2. Prova unilateral per la dreta $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$ $R = \left\{T : T \geq {}_{1-\alpha}t_{n-1}\right\}$

3. Prova unilateral per l'esquerra $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$ $R = \left\{T : T \leq -{}_{1-\alpha}t_{n-1}\right\}$

Exemple 2

Estem comprovant el funcionament d'una màquina que envasa fruita en almívar. Se suposa que els pots han de tenir 1 kg, però quan es desajusta sospitem que aquest pes augmenta perquè es veu que ix més fruita de la que estipulem, ja que s'afluixen els engranatges.

Per a comprovar-ho triem al llarg de la jornada 20 pots per a controlar-ne els pesos. Hem obtingut $\overline{x} = 995$ g i $S_{n-1} = 5{,}130$. Farem una prova unilateral per la dreta amb un nivell de significació de l'1%, plantejarem les hipòtesis nul·la i alternativa, i la regió crítica:

$$\begin{cases} H_0: \mu = 1.000 \\ H_1: \mu > 1.000 \end{cases} \quad R = \left\{T : T \geq {}_{1-\alpha}t_{n-1}\right\}.$$

Calculem l'estadístic amb les dades mostrals i poblacionals de l'enunciat:

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{995 - 1000}{\frac{5{,}130}{\sqrt{20}}} = \frac{-5}{1{,}1471} = -4{,}3588.$$

Per a trobar la regió crítica, calcularem el valor d'$1 - \alpha$ i el percentil corresponent de la variable t de Student amb 19 graus de llibertat:

$$\alpha = 0{,}01 \quad \to \quad 1 - \alpha = 0{,}99.$$

Per a trobar el valor de ${}_{1-\alpha}t_{n-1}$, calcularem el valor de x que faça complir que $P(T \leq x) = 0{,}9$, amb ajuda de les taules o d'un programa informàtic. Així ${}_{1-\alpha}t_{n-1} = {}_{0{,}99}t_{19} = 2{,}54$.

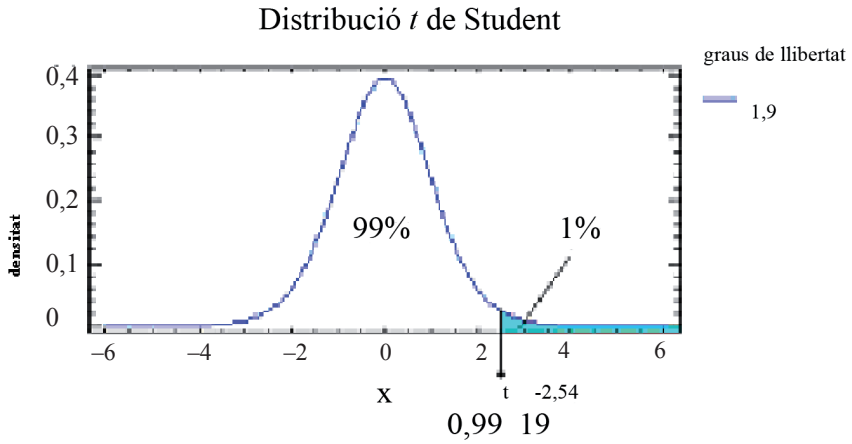


Figura 3. *Prova unilateral per la dreta*

$$R = R = \{T : T \geq {}_{1-\alpha}t_{n-1}\} = \{T : T \geq 2{,}54\} = (2{,}54, +\infty)$$

Podem comprovar que el valor del nostre estadístic $T = -4{,}3588 \notin R$ i això ens permet donar per veritable la hipòtesi nul·la i acceptar que la màquina funciona correctament. Gràficament es pot observar en la figura 3.

10.4.2. Altres contrastos d'hipòtesi

Contrast d'hipòtesi per a la proporció d'una població de Bernoulli

Estadístic $\mathrm{T} = \dfrac{\mathrm{p} - p}{\sqrt{\dfrac{pq}{n}}}$, que es distribueix $T \rightarrow N(0, 1)$. Denotarem per p la proporció en la població i per **p** la proporció en la mostra, de grandària n.

1. Prova bilateral $\begin{cases} H_o: \mathrm{p} = \mathrm{p}_o \\ H_1: \mathrm{p} \neq \mathrm{p}_o \end{cases}$ $\quad \mathrm{R} = \left\{\mathrm{T}: |\mathrm{T}| \geq z_{1-\frac{\alpha}{2}}\right\}$

2. Prova unilateral per la dreta $\begin{cases} H_o: \mathrm{p} = \mathrm{p}_o \\ H_1: \mathrm{p} > \mathrm{p}_o \end{cases}$ $\quad \mathrm{R} = \left\{\mathrm{T}: \mathrm{T} \geq -z_{1-\alpha}\right\}$

3. Prova unilateral per l'esquerra $\begin{cases} H_o: \mathrm{p} = \mathrm{p}_o \\ H_1: \mathrm{p} < \mathrm{p}_o \end{cases}$ $\quad \mathrm{R} = \left\{\mathrm{T}: \mathrm{T} \leq -z_{1-\alpha}\right\}$

Exemple 3

En aquest cas volem acabar l'exemple 2 amb el que hem començat el tema. Si el dau és correcte, la proporció de vegades que ix el 5 és $p = \frac{1}{6} = 0{,}167$, mentre que amb els nostre dau, hem fet 100 tirades i la proporció en la mostra és **p** = 0,25. Ens plantegem si el dau està trucat.

Plantejarem les hipòtesis nul·la i alternativa d'una prova bilateral i determinarem la regió crítica amb un nivell de significació del 5%.

$$\begin{cases} H_o: \text{p} = 0{,}167 \\ H_1: \text{p} \neq 0{,}167 \end{cases} \qquad \text{R} = \left\{\text{T: T} \geq -z_{1-\frac{\alpha}{2}}\right\}.$$

Calculem l'estadístic amb les dades mostrals i poblacionals de l'enunciat:

$$T = \frac{\text{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0{,}167 - 0{,}25}{\sqrt{\frac{0{,}25 \cdot 0{,}75}{100}}} = \frac{-0{,}083}{0{,}0433} = -1{,}9169.$$

Per a trobar la regió crítica, calcularem el valor d'$(1 - \alpha)$ i el percentil corresponent de la variable Z normal tipificada:

$$\alpha = 0{,}05 \quad \rightarrow \quad \frac{\alpha}{2} = 0{,}025 \quad \rightarrow \quad 1 - \frac{\alpha}{2} = 0{,}975.$$

Per a trobar el valor de $z_{1-\alpha}$, calcularem el valor de *x* que faça complir que $P\,(Z \leq x) = 0{,}975$, amb ajuda de les taules o d'un programa informàtic. Així, $z_{1-\frac{\alpha}{2}} = z_{0{,}975} = 1{,}96$.

$$R = \left\{T: |T| \geq z_{1-\frac{\alpha}{2}}\right\} = \left\{T: |T| \geq z_{0{,}975}\right\} = \left\{T: |T| \geq 1{,}96\right\} = (-\infty, -1{,}96) \cup (1{,}96, +\infty).$$

Podem comprovar que $Z = -1{,}9169 \notin R$, per tant no rebutjarem la hipòtesi nul·la i podrem assignar la diferència entre la proporció teòrica i la mostral a l'atzar. Així doncs, suposarem que el dau és correcte.

Contrast d'hipòtesi per a la variància d'una població normal

Estadístic $T = \dfrac{(n-1)S^2}{\sigma^2}$

Si denotem per $S^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ la quasivariància d'una mostra aleatòria de grandària *n*, extreta d'una població normal de variància σ^2, definirem com a estadístic la variable *T* com $T = \dfrac{(n-1)S^2}{\sigma^2} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2}$, que es distribueix com una variable khi quadrat amb $n - 1$ graus de llibertat, és a dir, $T \to \chi^2_{n-1}$.

1. Prova bilateral $\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 \neq \sigma_0^2 \end{cases}$ $R = \{T: T \leq {}_{\frac{\alpha}{2}}\chi^2_{n-1} \text{ o } T \geq {}_{1-\frac{\alpha}{2}}\chi^2_{n-1}\}$

2. Prova unilateral per la dreta $\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 > \sigma_0^2 \end{cases}$ $R = \{T: T \geq {}_{1-\alpha}\chi^2_{n-1}\}$

3. Prova unilateral per l'esquerra $\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 < \sigma_0^2 \end{cases}$ $R = \{T: T \geq {}_{\alpha}\chi^2_{n-1}\}$

Exemple 4

Tornarem a treballar sobre l'exemple 2 dels apartats anteriors i farem una estimació de la variància dels pesos de la màquina envasadora de pots de fruita. Considerarem que aquesta funciona correctament si la desviació típica del pesos és de 5 g, és a dir, la variància és 25.

Agafarem una mostra de 20 pots i prendrem els valors mostrals següents: la grandària de la mostra $n = 20$, la mitjana aritmètica dels pesos $\bar{x} = 995$ g i la seua desviació típica S = 5 g.

Necessitem calcular la quasivariància de la mostra i la podem obtenir a partir de la variància de la mostra:

$$S^2_{n-1} = \frac{n}{n-1}S^2 = \frac{20}{19}25 = 26{,}316.$$

Realitzarem una prova bilateral per a la variància σ^2 amb un nivell de significació del 5%, i la regió crítica R és:

$$\begin{cases} H_0: \sigma^2 = 25 \\ H_1: \sigma^2 \neq 25 \end{cases} \quad R = \{T: T \leq {}_{\frac{\alpha}{2}}\chi^2_{n-1} \ \text{ó} \ T \geq {}_{1-\frac{\alpha}{2}}\chi^2_{n-1}\}.$$

Utilitzarem l'expressió de l'estadístic anterior per a substituir els valors de la població i de la mostra:

$$T = \frac{(n-1)S^2}{\sigma^2} = \frac{19 \cdot 26{,}316}{25} = 20.$$

En aquest cas el nivell de significació escollit és del 5%, per la qual cosa:

$$1-\alpha = 0{,}95 \ \rightarrow \ \alpha = 0{,}05 \ \rightarrow \ \frac{\alpha}{2} = 0{,}025 \ \rightarrow \ 1-\frac{\alpha}{2} = 0{,}975.$$

Per a determinar la regió crítica necessitem buscar els valors de la variable ${}_{1-\frac{\alpha}{2}}\chi^2_{n-1} = {}_{0{,}975}\chi^2_{19}$ i ${}_{\frac{\alpha}{2}}\chi^2_{n-1} = {}_{0{,}025}\chi^2_{19}$ en la funció acumulada de la distribució khi quadrat amb 19 graus de llibertat, ja que com que la variable no és simètrica obtenim resultats diferents en valor absolut per a cada probabilitat.

En el nostre cas obtenim ${}_{0{,}975}X^2_{19} = 32{,}8523$ i que ${}_{0{,}025}X^2_{19} = 8{,}90652$. Els substituïm en la fórmula de la regió crítica i tenim:

$$R = \left\{T: T \leq {}_{\frac{\alpha}{2}}\chi^2_{n-1} \ \text{o} \ T \geq {}_{1-\frac{\alpha}{2}}\chi^2_{n-1}\right\} = \{T: T \leq 8{,}91 \ o \ T \geq 32{,}85\} = (-\infty; 8{,}91) \cup (32{,}85; +\infty).$$

És clar que el valor del nostre paràmetre està fora d'aquesta regió i, per tant, donem com a vàlida la hipòtesi nul·la: considerarem que la màquina funciona correctament, perquè el valor de la variància extret de la mostra no contradiu l'afirmació inicial que la variància dels pesos dels pots és de 25 g.

10.5. Contrastos d'hipòtesi per als paràmetres de dues poblacions

En l'apartat anterior hem vist els contrastos d'hipòtesi dels paràmetres d'una població per decidir si aquests tenen uns valors determinats a priori.

En aquest apartat farem els contrastos sobre els paràmetres de dues poblacions. Per a refutar o confirmar la hipòtesi de partida necessitarem les dades de dues mostres diferents extretes cadascuna de les respectives poblacions, ja que volem inferir conclusions sobre els valors dels seus paràmetres poblacionals.

Recordem que en aquests casos no podem decidir sobre els valors respectius dels paràmetres de les poblacions, sinó que podem comparar-los per diferència (en el cas de mitjanes i proporcions) o per quocient (en el cas de les variàncies), i concloure quin dels dos és més alt i en quina mesura.

Podem recordar l'exemple de la tercera situació presentada en la introducció d'aquest tema:

Suposem dues empreses que produeixen piles. Al llarg dels anys aquestes piles han tingut una durada similar, però en l'actualitat la segona empresa afirma que les seues duren més perquè ha fet una millora en la producció. Podem creure l'afirmació?

Amb el contrast d'hipòtesi podem fer-nos plantejaments sobre les mitjanes de la duració de les piles en cada empresa, del tipus:

- Duraran el mateix nombre d'hores les piles de totes dues empreses? És a dir, són iguals les mitjanes de A i B?
- Duraran les piles de l'empresa B més que les de A, tal com es presenta? És a dir, la mitjana de B és més alta que la de A?
- Duraran les piles de l'empresa B 5 hores més que les de l'empresa A? És a dir, la diferència de les mitjanes de A i B és de 5 hores?

Per a respondre a aquestes qüestions, caldrà triar una mostra representativa de cadascuna de les produccions de les empreses A i B de grandària n i m, les quals, en principi, poden ser diferents. Calcularem les mesures que necessitem en cadascuna i analitzarem si aquestes mesures ens fan dubtar, en termes de probabilitat, de la versemblança de les nostres hipòtesis inicials, o no.

Recordem que $X_1, X_2, \ldots X_n$ és una mostra aleatòria de grandària n extreta d'una població, i $Y_1, Y_2, \ldots, Y_m$ la mostra aleatòria de grandària m extreta d'una altra població.

Mantenim el paral·lelisme amb els valors de les mitjanes μ_x, μ_y per a cadascuna de les poblacions, i els símbols σ_x^2 i σ_Y^2 per a denotar-ne les variàncies.

Els valors de les mitjanes mostrals són $\overline{X}$ i $\overline{Y}$. Les quasivariàncies mostrals són $S^2_{x_{n-1}}$ i $S^2_{Y_{n-1}}$.

En el cas d'una població de Bernoulli, denotem per p_x, p_y les proporcions obtingudes a partir dels valors de les mostres, i per p_x, p_y les proporcions poblacionals que cal estimar de les poblacions respectives.

Quant a les notacions de les variables, i les expressions dels percentils, són els que ja hem utilitzat en els temes anteriors.

La resta d'aspectes respecte al disseny d'un test són iguals als que es van tractar en l'apartat anterior i anirem comentant-los amb l'exemple de cada cas.

10.5.1. Contrast d'hipòtesi per a la diferència de mitjanes

Totes les consideracions que férem sobre la comparació de mitjanes en l'estimació per intervals en el tema anterior, pel que fa a la possibilitat de treballar amb mostres independents o emparellades, es tornarien a plantejar en aquest apartat. Per la qual cosa també ací presentem els tres casos que hem desenvolupat en els tres temes d'inferència:

- Diferència de mitjanes amb mostres independents i variàncies poblacionals conegudes.
- Diferència de mitjanes amb mostres independents i variàncies poblacionals desconegudes, però que podem suposar iguals.
- Diferència de mitjanes amb mostres dependents i/o emparellades.

És interessant que vegem com plantejaríem les hipòtesis nul·la i alternativa en els casos de la introducció.

- Duraran el mateix nombre d'hores les piles de totes dues empreses? És a dir, són iguals les mitjanes de *A* i *B*? $\begin{cases} H_0 : \mu_X - \mu_Y = 0 \\ H_1 : \mu_X - \mu_Y \neq 0 \end{cases}$

- Duraran les piles de l'empresa *B* més que les de *A* tal com es presenta? És a dir, la mitjana de *B* és més gran que la de *A*? $\begin{cases} H_0 : \mu_X - \mu_Y = 0 \\ H_1 : \mu_X - \mu_Y < 0 \end{cases}$

- Duraran les piles de l'empresa *B* 5 hores més que les de l'empresa *A*? És a dir, la diferència de les mitjanes de *A* i *B* és de 5 hores? $\begin{cases} H_0 : \mu_X - \mu_Y = -5 \\ H_1 : \mu_X - \mu_Y \neq -5 \end{cases}$

Contrast d'hipòtesi per a la diferència de mitjanes de poblacions independents amb variàncies poblacionals conegudes

Estadístic $T = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}}$ que es distribueix $T \to N(0, 1)$.

1. Prova unilateral $\begin{cases} H_0: \mu_X - \mu_Y = d_0 \\ H_1: \mu_X - \mu_Y \neq d_0 \end{cases}$ $R = \{T: |T| \geq z_{1-\frac{\alpha}{2}}\}$

2. Prova unilateral per la dreta $\begin{cases} H_0: \mu_X - \mu_Y = d_0 \\ H_1: \mu_X - \mu_Y > d_0 \end{cases}$ $R = \{T: |T| \geq z_{1-\alpha}\}$

3. Prova unilateral per l'esquerra $\begin{cases} H_0: \mu_X - \mu_Y = d_0 \\ H_1: \mu_X - \mu_Y < d_0 \end{cases}$ $R = \{T: |T| \leq -z_{1-\alpha}\}$

Exemple 5

Volem comparar l'eficiència de dues empreses de missatgeria internacional, atenent el nombre d'hores que tardem a rebre'n les remeses. Creiem que el funcionament de totes dues és semblant si en comparem les mitjanes i coneixem les desviacions típiques de l'empresa A (25 hores) i de la de B (30 hores.) Ambdues es comporten com a variables normals.

Per a confirmar la nostra creença, anotem en 10 enviaments de l'empresa A, un temps mitjà de 80 hores, mentre que en la mostra dels 15 enviaments de l'empresa B el temps mitjà és de 75 hores.

Dissenyem una prova bilateral. Definim l'estadístic T i decidim treballar amb un nivell de significació de l'1% per a decidir la regió crítica. Així:

$$\begin{cases} H_0: \mu_X - \mu_Y = 0 \\ H_1: \mu_X - \mu_Y \neq 0 \end{cases} \quad R = \{T: T \geq z_{1-\frac{\alpha}{2}}\}$$

Calculem l'estadístic amb les dades mostrals i poblacionals de l'enunciat, del qual coneixem que es distribueix com una variable normal tipificada.

Per a la població A, podem anotar el paràmetre de la població $\sigma_x^2 = 25^2 = 625$, i amb les dades de la mostra $n = 10$, $\overline{x} = 80$.

Per a la població B, podem anotar el paràmetre de la població $\sigma_Y^2 = 30^2 = 900$, i amb les dades de la mostra $m = 15$, $\overline{y} = 75$.

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} = \frac{(80 - 75) - 0}{\sqrt{\frac{625}{10} + \frac{900}{15}}} = \frac{5}{11{,}0680} = 0{,}4518.$$

Per a trobar la regió crítica amb un nivell de significació de l'1%, calcularem el valor d'$1 - \frac{\alpha}{2}$ i el percentil corresponent de la variable *Z* normal tipificada:

$$\alpha = 0{,}01 \quad \rightarrow \quad \frac{\alpha}{2} = 0{,}005 \quad \rightarrow \quad 1 - \frac{\alpha}{2} = 0{,}995.$$

i en la taula de la distribució acumulada de la variable normal tipificada, trobem el valor de la variable *x* que faça complir que $P(Z \leq x) = 0{,}995$. Aquest valor és *x* = = 2,57583, per la qual cosa, utilitzant la notació d'aquest tema, $z_{1-\frac{\alpha}{2}} = z_{0{,}995} = 2{,}58$, que substituïm en l'expressió de *R*:

$$R = \left\{T : |T| \geq z_{1-\frac{\alpha}{2}}\right\} = \left\{T : |T| \geq 2{,}58\right\} = (-\infty; -2{,}58) \cup (2{,}58; +\infty).$$

$T = 0{,}4518 \notin R$, per tant no rebutgem la hipòtesi nul·la i podem donar per bona l'afirmació que totes dues empreses tenen les mitjanes iguals i ens semblen igual d'eficaces.

Contrast d'hipòtesi per a la diferència de mitjanes de poblacions independents amb variàncies poblacionals desconegudes però iguals

Estadístic: Si denotem per $S^2_{X_{m-1}} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$ la quasivariància de la mostra aleatòria de grandària *n* extreta de la primera població i per $S^2_{Y_{m-1}} = \frac{\sum_{i=1}^{m} (y_i - \overline{y})^2}{m-1}$ la quasivariància de la mostra de grandària *m* de la segona població, definim la variable $T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{(n-1)S^2_{x_{n-1}} + (m-1)S^2_{Y_{m-1}}}{n+m-2}} \sqrt{\frac{1}{n} + \frac{1}{m}}}$, que és l'estadístic que es distribueix segons el model d'una variable *t* de Student amb *n* + *m* – 2 graus de llibertat. És a dir, $T \rightarrow t_{n+m-2}$:

1. Prova unilateral $\begin{cases} H_0 : \mu_X - \mu_Y = d_0 \\ H_1 : \mu_X - \mu_Y \neq d_0 \end{cases} \quad R = \{T : T \geq {}_{1-\frac{\alpha}{2}}t_{n+m-2}\}$

2. Prova unilateral per la dreta $\begin{cases} H_0 : \mu_X - \mu_Y = d_0 \\ H_1 : \mu_X - \mu_Y > d_0 \end{cases} \quad R = \{T : T \geq {}_{1-\alpha}t_{n+m-2}\}$

3. Prova unilateral per l'esquerra $\begin{cases} H_0 : \mu_X - \mu_Y = d_0 \\ H_1 : \mu_X - \mu_Y < d_0 \end{cases} \quad R = \{T : T \leq -{}_{1-\alpha}t_{n+m-2}\}$

Aplicarem aquesta fórmula quan no coneguem les variàncies poblacionals σ_x^2 ni σ_Y^2, però puguem afirmar que són iguals. Si no fóra el cas, podríem fer-hi un contrast per a estimar la igualtat entre totes dues variàncies, com es veurà un poc més endavant en aquest mateix apartat, i si la interpretació ens permet estimar que són iguals continuarem calculant aquesta diferència de mitjanes.

Exemple 6

En una mesura de control de qualitat en la fabricació d'unes peces, volem comparar si dos processos de producció són equivalents i mantenen els mateixos estàndards de qualitat. Considerarem que les variàncies de totes dues poblacions són iguals. Sabem que l'empresa B ha millorat i se suposa que aquesta es pot concretar almenys en un punt més en la mitjana de les puntuacions del test que utilitzem com a indicador.

Per a realitzar el treball, agafem unes quantes peces de cada línia i amb aquestes dades, suposant que $\sigma_x^2 = \sigma_Y^2$, obtenim aquests resultats.

De la mostra de la població X, anotarem $n = 10$, $\bar{x} = 7{,}9$ i $S^2_{x_{n-1}} = 6{,}77$.

De la mostra de la població Y, anotarem $m = 12$, $\bar{y} = 9{,}75$ i $S^2_{Y_{n-1}} = 6{,}39$.

Dissenyarem una prova unilateral per l'esquerra, amb un nivell de significació del 10%.

$$\begin{cases} H_0 : \mu_X - \mu_Y = -1 \\ H_1 : \mu_X - \mu_Y < -1 \end{cases} \qquad R = \{T : T \leq -{}_{1-\alpha}t_{n+m-2}\}$$

Calculem l'estadístic amb les dades mostrals i poblacionals de l'enunciat, del qual sabem que es distribueix com una variable *t* de Student amb 20 graus de llibertat:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{(n-1)S^2_{x_{n-1}} + (m-1)S^2_{Y_{m-1}}}{n+m-2}}\sqrt{\frac{1}{n}+\frac{1}{m}}} = \frac{(7{,}9 - 9{,}75) - (-1)}{\sqrt{\frac{9 \cdot 6{,}77 + 11 \cdot 6{,}39}{20}}\sqrt{\frac{1}{10}+\frac{1}{12}}} = \frac{-0{,}85}{1{,}0967} = -0{,}7750.$$

Per a conèixer la regió crítica, cal esbrinar els valors dels extrems d'aquesta. Cal trobar el valor de $_{1-\alpha}t_{n+m-2}$ utilitzant un nivell de significació del 10%:

$$\alpha = 0{,}10 \quad \rightarrow \quad 1 - \alpha = 0{,}90.$$

Així, cal calcular en les taules de la distribució acumulada de la variable *t* de Student amb 20 graus de llibertat, el valor de la variable $_{1-\alpha}t_{n+m-2} = {}_{0{,}90}t_{20} = 1{,}3253$, i ara substituïm aquest valor en la fórmula de la regió crítica R:

$$R = \{T : T \leq -{}_{1-\alpha}t_{n+m-2}\} = \{T : T \leq -{}_{0{,}90}t_{29}\} = (-\infty, -1{,}3253).$$

Com podem veure, $T = -0{,}7750 \notin R$, per la qual cosa no podem refutar la hipòtesi nul·la, i direm que, efectivament, hi ha un punt de diferència a favor de l'empresa B en la mitjana de les puntuacions de l'indicador de qualitat, però la informació de les mostres no permet inferir que la diferència serà més gran.

Contrast d'hipòtesi per a la diferència de mitjanes de dues poblacions amb mostres relacionades

Aquest contrast, l'apliquem, com ja es va explicar en els apartats dels temes anteriors d'estimació, quan volem comparar la mitjana de dues poblacions que estan relacionades i les mostres estan formades per parells (x_i, y_i).

Estadístic: $T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\frac{S_{D_{n-1}}}{\sqrt{n}}}$, del qual sabem que té una distribució que segueix el model d'una variable *t* de Student amb $n - 1$ graus de llibertat. És a dir, $T \rightarrow t_{n-1}$.

Recordem que en aquest cas cal definir prèviament la variable $D = X - Y$, per a la qual denotarem per $S_{D_{n-1}}$ la quasidesviació típica mostral, on $d_i = x_i - y_i$. La grandària de totes dues mostres necessàriament coincideix i serà *n*.

1. Prova unilateral $\begin{cases} H_0: \mu_X - \mu_Y = d_0 \\ H_1: \mu_X - \mu_Y \neq d_0 \end{cases} \quad R = \{T: T \geq {}_{1-\frac{\alpha}{2}}t_{n-1}\}$

2. Prova unilateral per la dreta $\begin{cases} H_0: \mu_X - \mu_Y = d_0 \\ H_1: \mu_X - \mu_Y > d_0 \end{cases} \quad R = \{T: T \geq {}_{1-\alpha}t_{n-1}\}$

3. Prova unilateral per l'esquerra $\begin{cases} H_0: \mu_X - \mu_Y = d_0 \\ H_1: \mu_X - \mu_Y < d_0 \end{cases} \quad R = \{T: T \leq -{}_{1-\alpha}t_{n-1}\}$

Exemple 7

Per a millorar el grau de satisfacció dels clients del banc A s'ha plantejat eliminar les comissions que cobraven als clients per alguns serveis. Suposem que aquesta mesura ha reportat una millora en el grau de satisfacció dels clients.

Per a avaluar l'eficàcia d'aquesta decisió, s'ha passat a vuit clients una enquesta dissenyada per a esbrinar el grau mitjà de satisfacció abans i després de l'eliminació de les comissions en una escala de 0 a 3.

Amb les respostes s'obtenen els resultats següents de les mostres de les variables X (abans) i Y (després): $n = 8$, $\bar{x} = 1{,}45$, $\bar{y} = 1{,}6625$ i amb les dades de la variable D, calculem $S^2_{D_{n-1}} = 0{,}0327$.

Dissenyem una prova unilateral per l'esquerra (ens sembla obvi descartar que puga empitjorar com a conseqüència de la mesura) amb un grau de significació del 5%:

$$\begin{cases} H_0: \mu_X - \mu_Y = 0 \\ H_1: \mu_X - \mu_Y < 0 \end{cases} \quad R = \{T: T \leq -{}_{1-\alpha}t_{n-1}\}$$

Calculem l'estadístic amb les dades mostrals i poblacionals de l'enunciat, del qual sabem que es distribueix com una variable t de Student amb 7 graus de llibertat:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\frac{S_{D_{n-1}}}{\sqrt{n}}} = \frac{(1{,}45 - 1{,}6625) - 0}{\frac{0{,}0327}{\sqrt{8}}} = \frac{-0{,}2125}{0{,}0116} = -18{,}3190.$$

Per a conèixer la regió crítica, cal esbrinar els valors dels extrems d'aquesta. Cal trobar el valor de ${}_{1-\alpha}t_{n-1}$ utilitzant un nivell de significació del 5%:

$$\alpha = 0{,}05 \quad \rightarrow \quad 1 - \alpha = 0{,}95.$$

Així, cal calcular en les taules de la distribució acumulada de la variable *t* de Student amb 7 graus de llibertat, el valor de la variable $_{1-\alpha}t_{n-1}=_{0,95}t_7=1,8946$ i ara substituïm aquest valor en la fórmula de la regió crítica R:

$$R=\{T:T\le -_{1-\alpha}t_{n-1}\}=\{T:T\le -_{0,95}t_7\}=(-\infty,-1,8946).$$

És evident que $T\in R$. La informació de les mostres contradiu la hipòtesi inicial, i ens porta a rebutjar-la; com a conseqüència, podem dir que la mesura sí que ha reportat una millora en el grau de satisfacció dels clients, tal com indicava la nostra hipòtesi alternativa.

10.5.2. Altres contrastos d'hipòtesi

Contrast d'hipòtesi per a la diferència de proporcions de dues poblacions de Bernoulli

Estadístic: $T=\dfrac{(p_x-p_Y)-(\mathrm{p_X}-\mathrm{p_Y})}{\sqrt{\dfrac{p_x(1-p_x)}{n}+\dfrac{p_Y(1-p_Y)}{m}}}$ i sabem que T es distribueix com una normal tipificada.

Recordem que denotem per $\mathbf{p}_x$, $\mathbf{p}_y$ les proporcions obtingudes a partir dels valors de les mostres i per p_x, p_y les proporcions poblacionals de la diferència de les quals farem el contrast.

1. Prova bilateral $\begin{cases}H_0: p_X-p_Y=d_0\\ H_1: p_X-p_Y\neq d_0\end{cases}$ $R=\{T:|T|\ge z_{1-\frac{\alpha}{2}}\}$
2. Prova unilateral per la dreta $\begin{cases}H_0: p_X-p_Y=d_0\\ H_1: p_X-p_Y>d_0\end{cases}$ $R=\{T:T\ge z_{1-\alpha}\}$
3. Prova unilateral per l'esquerra $\begin{cases}H_0: p_X-p_Y=d_0\\ H_1: p_X-p_Y<d_0\end{cases}$ $R=\{T:T\le -z_{1-\alpha}\}$

Exemple 8

Realitzant un control de qualitat en la maquinària d'una empresa, volem corroborar si les dues màquines que tenim són igual d'eficients considerant la proporció de peces defectuoses que ixen de cadascuna en el procés d'elaboració. Per a realitzar el treball seleccionem aleatòriament una mostra de 200 peces de la màquina A, de les quals 15 són defectuoses i 250 peces de la màquina B, de les quals 16 són defectuoses.

Anotem aquestes dades mostrals $n = 200 \qquad p_X = \frac{15}{200} = 0{,}075$

$$m = 250 \qquad p_Y = \frac{16}{250} = 0{,}064.$$

Dissenyem una prova bilateral amb un nivell de significació del 5% per a la diferència de proporcions:

$$\begin{cases} H_0 : p_X - p_Y = 0 \\ H_1 : p_X - p_Y \neq 0 \end{cases} \qquad R = \{T: |T| \geq z_{1-\frac{\alpha}{2}}\}.$$

I substituïm les dades en l'expressió de l'estadístic:

$$T = \frac{(p_x - p_Y) - (p_X - p_Y)}{\sqrt{\frac{p_x(1-p_x)}{n} + \frac{p_Y(1-p_Y)}{m}}} = \frac{(0{,}075 - 0{,}064) - 0}{\sqrt{\frac{0{,}075 \cdot 0{,}925}{200} + \frac{0{,}064 \cdot 0{,}936}{250}}} = \frac{0{,}011}{0{,}0242} = 0{,}4545.$$

Com que treballarem amb un nivell de confiança del 5%, podem calcular el valor d'$1-\frac{\alpha}{2}$, ja que $\alpha = 0{,}05 \rightarrow \frac{\alpha}{2} = 0{,}025 \rightarrow 1-\frac{\alpha}{2} = 0{,}975$, i en la taula de la distribució acumulada de la variable normal tipificada, busquem el valor de la variable x que faça complir que $P(Z \leq x) = 0{,}975$. Aquest valor és $x = 1{,}96$, per la qual cosa $z_{1-\frac{\alpha}{2}} = 1{,}96$. Substituïm aquest valor en l'expressió de la regió crítica i obtenim:

$$R = \{T: |T| \geq z_{1-\frac{\alpha}{2}}\} = R = \{T: |T| \geq 1{,}96\} = (-\infty, -1{,}96) \cup (1{,}96; +\infty).$$

És clar que $T = 0{,}45 \notin R$. La conclusió que podem extraure d'aquest resultat és que treballant amb un nivell de significació del 5%, la diferència entre les proporcions mostrals que hem obtingut no és significativa, i podem estimar que la proporció de peces defectuoses en la producció de totes dues màquines és la mateixa.

Contrast d'hipòtesi per al quocient de dues variàncies de dues poblacions normals

Estadístic: $T = \dfrac{\dfrac{(n-1)S^2_{X_{n-1}}}{\sigma^2_X \cdot n}}{\dfrac{(m-1)S^2_{Y_{n-1}}}{\sigma^2_Y \cdot m}}$, del qual ja sabem que es distribueix com una variable F de Fisher-Snedecor amb $n-1$ i $m-1$ graus de llibertat, respectivament.

1. Prova bilateral $\begin{cases} H_0: \dfrac{\sigma^2_X}{\sigma^2_Y} = \sigma^2_0 \\ H_1: \dfrac{\sigma^2_X}{\sigma^2_Y} \neq \sigma^2_0 \end{cases}$ $R = \{T: T \leq {}_{\frac{\alpha}{2}}f_{n-1,m-1} \text{ o } T \geq {}_{1-\frac{\alpha}{2}}f_{n-1,\ m-1}\}$

2. Prova unilateral per la dreta $\begin{cases} H_0: \dfrac{\sigma^2_X}{\sigma^2_Y} = \sigma^2_0 \\ H_1: \dfrac{\sigma^2_X}{\sigma^2_Y} > \sigma^2_0 \end{cases}$ $R = \{T: T \geq {}_{1-\alpha}f_{n-1,\ m-1}\}$

3. Prova unilateral per l'esquerra $\begin{cases} H_0: \dfrac{\sigma^2_X}{\sigma^2_Y} = \sigma^2_0 \\ H_1: \dfrac{\sigma^2_X}{\sigma^2_Y} < \sigma^2_0 \end{cases}$ $R = \{T: T \leq {}_{\alpha}f_{n-1,\ m-1}\}$

Exemple 9

Per a estudiar aquest contrast, prendrem com a exemple les dades de l'exemple 6, en el qual per a comparar les mitjanes de dues poblacions havíem utilitzat les dues mostres que presentem a continuació.

Recordem que en aquell apartat ja vam comentar que es tractava de dues mostres, les poblacions de les quals necessitem pressuposar que tenen les variàncies iguals. Si aquesta circumstància no la coneixem a priori per treballs anteriors, caldrà començar fent el treball que presentem a continuació. En cas que la inferència ens permeta estimar que són iguals, podrem portar a terme el treball que ja vàrem fer en l'exemple 6.

D'altra banda, cal recordar també que es tractava d'unes quantes peces agafades de dues línies de producció i que estaven classificades amb l'ajuda d'un índex de qualitat que resumeix la informació de diversos indicadors.

Amb aquesta informació volem estimar, mitjançant un contrast d'hipòtesi, si la variabilitat en la qualitat de totes dues línies de producció és la mateixa. Entenem per *variabilitat* el valor de les variàncies, que és el paràmetre que utilitzarem com a indicador de la dispersió, ja que volem comprovar si els indicadors de qualitat permeten comprovar que estan igualment propers a la seua mitjana en tots dos processos.

Per a dur a terme els càlculs, identificarem els valors dels paràmetres de cadascuna de les mostres:

De la mostra de la població X, anotarem $n = 10$, $\bar{x} = 7{,}9$ i $S^2_{x_{n-1}} = 6{,}77$.

De la mostra de la població Y, anotarem $m = 12$, $\bar{y} = 9{,}75$ i $S^2_{Y_{n-1}} = 6{,}39$.

Dissenyarem una prova bilateral amb un nivell de significació del 5%.

$$\begin{cases} H_0: \dfrac{\sigma_X^2}{\sigma_Y^2} = 1 \\ H_1: \dfrac{\sigma_X^2}{\sigma_Y^2} \neq 1 \end{cases} \qquad R = \{T: T \leq {}_{\frac{\alpha}{2}}f_{n-1,m-1} \text{ o } T \geq {}_{1-\frac{\alpha}{2}}f_{n-1,\,m-1}\}$$

Comencem per substituir en l'estadístic els valors anteriors:

$$T = \frac{\dfrac{(n-1)S^2_{X_{n-1}}}{\sigma_X^2 \cdot n}}{\dfrac{(m-1)S^2_{Y_{n-1}}}{\sigma_Y^2 \cdot m}} = \frac{\sigma_Y^2 \cdot m(n-1)S^2_{X_{n-1}}}{\sigma_X^2 \cdot n(m-1)S^2_{Y_{n-1}}} = \frac{12 \cdot 9 \cdot 6{,}77}{10 \cdot 11 \cdot 6{,}39} = \frac{731{,}16}{702{,}9} = 1{,}0402.$$

Treballarem amb un nivell de significació del 5%, per la qual cosa:

$$\alpha = 0{,}05 \quad \rightarrow \quad \frac{\alpha}{2} = 0{,}025 \quad \rightarrow \quad 1 - \frac{\alpha}{2} = 0{,}975.$$

Buscarem en la taula de la distribució acumulada de la variable F de Fisher-Snedecor, els percentils que corresponen al valors de $\frac{\alpha}{2} = 0{,}025$ i $1 - \frac{\alpha}{2} = 0{,}975$. Cal comentar que si utilitzem taules, calcularem el primer valor en funció del segon, que sí que podrem trobar en les taules que solen publicar-se, aplicant-hi la propietat que expliquem a continuació de la funció de distribució d'aquesta variable *F* de Fisher-Snedecor.

Així, en les taules o amb un programa estadístic, trobem el valor del percentil $_{1-\frac{\alpha}{2}}f_{(n-1),(m-1)}=_{0,975}f_{9,1} = 3{,}5879$ i per a calcular l'altre valor farem ús de la propietat $_{\frac{\alpha}{2}}f_{(n-1),(m-1)} = \frac{1}{_{1-\frac{\alpha}{2}}f_{(m-1),(n-1)}}$, que aplicada al nostre cas seria $_{0,025}f_{9,1} =$

$$= \frac{1}{_{0,975}f_{1,9}} = \frac{1}{3{,}91207} = 0{,}255619.$$

Si substituïm tots aquests valors en l'expressió de la regió crítica obtindrem com a resultat que:

$$R = \left\{T{:}T \leq {}_{\frac{\alpha}{2}}f_{n-1,m-1} \text{ o } T \geq {}_{1-\frac{\alpha}{2}}f_{n-1,m-1}\right\} = \left\{T{:}T \leq {}_{0,025}f_{9,11} \text{ o } T \geq {}_{0,975}f_{9,11}\right\} =$$

$$= \{T{:}T \leq 0{,}2556 \text{ o } T \geq 3{,}5879\} = (-\infty, 0{,}26) \cup (3{,}59, +\infty).$$

Com que el valor de $T = 1{,}040 \notin R$ permet inferir que la hipòtesi nul·la és veritable i conseqüentment, les variàncies de les poblacions de les quals provenen les mostres són iguals, és a dir, els diferents productes de cadascuna de les línies de fabricació que comparem, presenten el mateix comportament en la dispersió dels valors de qualitat respecte a la mitjana de cadascuna d'aquelles.

Nota

Una vegada arribats a aquest punt del desenvolupament del tema, volem remarcar que la prova bilateral es pot abordar també des del tema d'intervals de confiança.

Si cal comprovar que un paràmetre de la població té un cert valor $\theta = \theta_0$ (hipòtesi nul·la) o $\theta \neq \theta_0$ (hipòtesi alternativa) amb un nivell de significació α, serà equivalent a veure si en l'interval de confiança corresponent al dit paràmetre, en el qual es compleix $P(\theta \in (a, b)) = 1-\alpha$ podem observar que $\theta_0 \in (a, b)$.

No es pot dir el mateix de les proves unilaterals per la dreta o per l'esquerra.

Si es comparen els exemples dels temes 9 i 10, es pot comprovar que majoritàriament hem utilitzat els mateixos exemples per a afavorir el plantejament paral·lel de les dues tècniques d'inferència. Convidem el lector/a a llegir i analitzar-los comparant tècnica i resultats.

10.6. Valor p

A hores d'ara ja coneixem que el resultat d'un contrast d'hipòtesi depèn del nivell de significació que triem per realitzar-lo. En algunes situacions el valor de l'estadístic queda prop dels valors frontera de la regió crítica, per la qual cosa resulta interessant conèixer quin seria el nivell de significació amb el qual aquest valor de l'estadístic ens faria rebutjar la hipòtesi nul·la.

Per això és raonable i freqüent adjuntar a la decisió de rebutjar H_0 el que direm *valor p*.

Definirem *valor p* com el mínim grau de significació amb el qual es pot rebutjar la hipòtesi nul·la amb el valor de l'estadístic T que hem calculat amb les dades del plantejament del test. És a dir, si α fóra més gran que el valor p rebutjaríem la hipòtesi de partida, i si fóra menor, l'acceptaríem.

Cal diferenciar els tres tipus de contrast:

1. Si és prova bilateral $p = 2\min\{P(T \geq t), P(T \leq t)\}$
2. Si és prova unilateral per la dreta $p = \{P(T \geq t)\}$
3. Si és prova unilateral per l'esquerra $p = \{P(T \leq t)\}$

Exemple 10

Dels resultats de les decisions que hem realitzat en els exemples desenvolupats al llarg del tema, podem triar l'exemple 3 per a aplicar el valor p, ja que el valor de l'estadístic que hi vam calcular estava molt proper als extrems dels intervals de la regió crítica.

Recordem l'exemple 3: si el dau és correcte, la proporció de vegades que ix el 5 és $p = \frac{1}{6} = 0{,}167$, mentre que amb el nostre dau, hem fet 100 tirades i la proporció en la mostra és $p = 0{,}25$. Ens plantegem si el dau està trucat.

Plantejarem les hipòtesis nul·la i alternativa d'un contrast bilateral $\begin{cases} H_0 : \mathrm{p} = 0{,}167, \\ H_1 : \mathrm{p} \neq 0{,}167. \end{cases}$

Calculem l'estadístic amb les dades mostrals i poblacionals de l'enunciat, sabent que aquest es distribueix com una normal tipificada:

$$T = \frac{\mathrm{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0{,}167 - 0{,}25}{\sqrt{\frac{0{,}25 \cdot 0{,}75}{100}}} = \frac{-0{,}083}{0{,}0433} = -1{,}9169.$$

Cal calcular, amb els valors de la funció de distribució de la variable Z, la probabilitat que $P(Z \leq -1{,}9169) = 0{,}0277$ per tractar-se d'una funció simètrica.

Així: $p = 2\min\{P(T \geq t), P(T \leq t)\} = 2 \cdot 0{,}0277 = 0{,}0554.$

Per la qual cosa, si treballem amb un nivell de significació del 5%, per ser inferior al valor p acceptaríem la hipòtesi nul·la (com passava en l'exemple 3), però per a valors superiors a 5,5% de nivell de significació, la hipòtesi nul·la hauria estat rebutjada i conclouríem que el dau està trucat.

10.7. Problemes proposats

En aquest epígraf es plantejaran un conjunt de problemes per a la resolució dels quals és necessari conèixer la teoria desenvolupada al llarg de la unitat.

Exercici 1

En un treball d'investigació social, hem comprovat que en una mostra de 823 adults, el 53% estaven d'acord amb les mesures de discriminació positiva per a la representació política de les dones en les institucions, el 37% hi estaven en desacord i el 10% no n'estaven segurs. Realitza un contrast amb un $\alpha = 0{,}05$, per a estimar si la majoria de la població està a favor d'aquestes mesures.

Exercici 2

Per a reforçar les estructures d'un pont en una autopista es coneix que en aquesta via es pot considerar una mitjana de 72 vehicles per hora en un tram de 25 km. Els enginyers sospiten que en el tram del nou pont, el volum de trànsit és més gran i per contrastar-ho de manera aleatòria mesuren en 50 hores diferents al llarg del mes, el nombre de vehicles que utilitzen aquest tram de 25 km que inclou el pont que cal millorar. S'hi obté una mitjana de 74,1 vehicles amb una desviació típica de 13,3, respectivament. Confirma aquestes mesures observades el valor de la mitjana que els havien proporcionat amb un nivell de significació del 10%?

Exercici 3

Per a estalviar en les despeses he decidit fer un estudi de les consumicions diàries en el dinar, al lloc de treball. He anotat les despeses dels darrers 20 dies laborals en aquest concepte i he obtingut una mitjana de 21 € amb una desviació típica de 17 €. Podem estimar mitjançant contrast d'hipòtesi que les meues despeses diàries són superiors a 10 € amb un nivell de significació del 5%?

Exercici 4

Per a donar per útil un model d'envasadora de malles de taronges de 8 kg hem decidit controlar-ne la dispersió respecte de l'etiquetatge. Per a això, triem a l'atzar 10 malles i n'anotem el pes (en kg).

7,96	7,90	7,98	8,01	7,97	7,96	8,03	8,02	8,04	8,02

Per a donar per vàlid el funcionament, la desviació ha de ser inferior a 0,1 kg. Permeten aquestes mesures donar per correcta aquesta màquina per un contrast d'hipòtesi, amb un nivell de significació del 5%?

Exercici 5

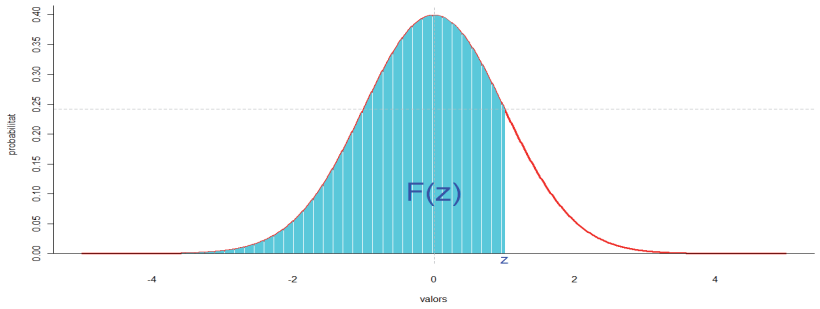
La nostra empresa sempre havia contractat l'agència A per a fer les campanyes publicitàries i d'estudis de mercat. El preu dels seus serveis es distribueix com un model normal de mitjana 820 € amb una desviació típica de 80 €. Una agència B ens ha estat donant ofertes molt temptadores per als darrers treballs, però no volem fer el canvi si les despeses no tenen un valor de la mitjana inferior. Per a això, una empresa independent ha estat fent un estudi i ha anotat 36 possibles serveis, i ha obtingut una mitjana de 790 €. Per a σ = 80 €, realitza un contrast d'hipòtesi per a aconsellar el canvi o no, amb un nivell de significació del 5%.

Exercici 6

Per justificar el llançament de les marques blanques com a mesura davant la crisi i per augmentar els beneficis d'una determinada empresa, fem un estudi per a una superfície comercial i preguntem als clients si coneixen el nom de la marca corresponent. Tan sols 47 dels 102 entrevistats la coneixen. Podem afirmar que menys de la meitat del clients la coneixen mitjançant un contrast d'hipòtesi, amb un nivell de significació del 5%?

TAULES ESTADÍSTIQUES

Taula 1. Distribució normal (0,1)

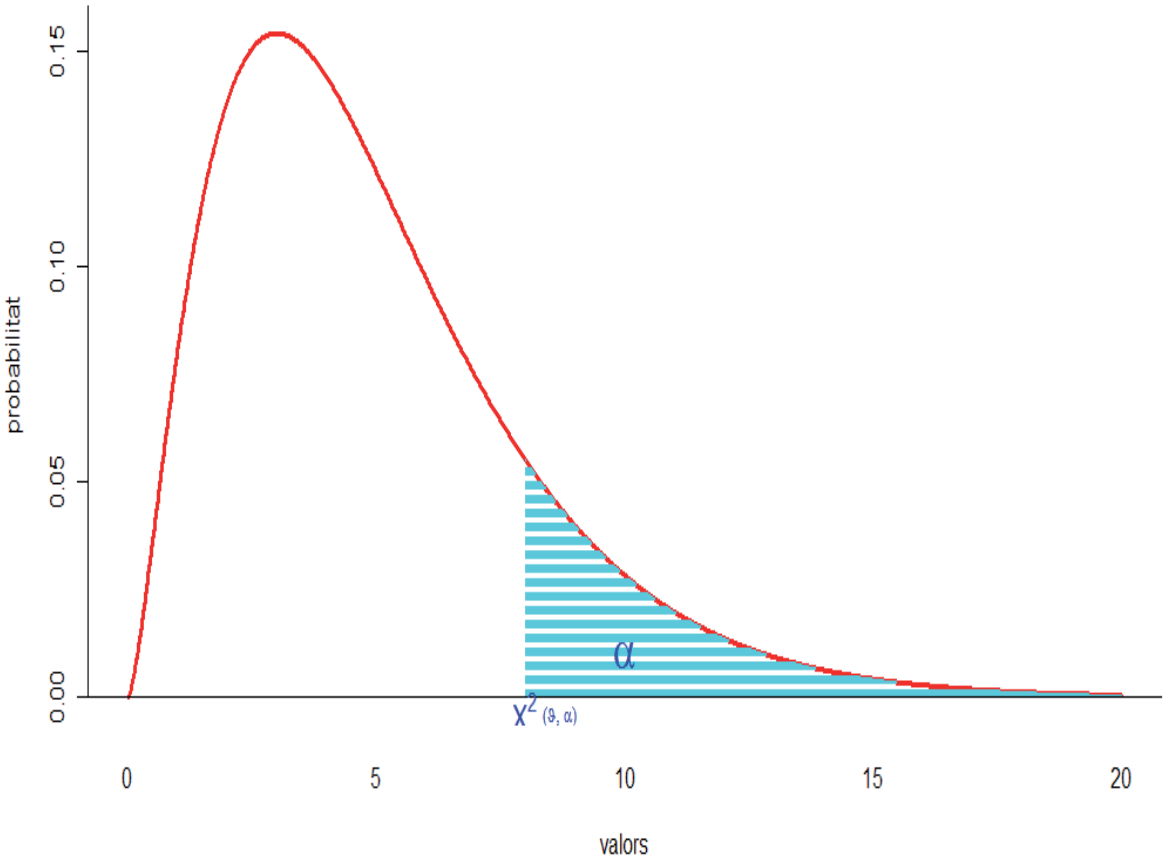


z	F(z)	z	F(z)	z	F(z)	z	F(z)	z	F(z)
0.00	0.5000	0.22	0.5871	0.44	0.6700	0.66	0.7454	0.88	0.8106
0.01	0.5040	0.23	0.5910	0.45	0.6736	0.67	0.7486	0.89	0.8133
0.02	0.5080	0.24	0.5948	0.46	0.6772	0.68	0.7517	0.90	0.8159
0.03	0.5120	0.25	0.5987	0.47	0.6808	0.69	0.7549	0.91	0.8186
0.04	0.5160	0.26	0.6026	0.48	0.6844	0.70	0.7580	0.92	0.8212
0.05	0.5199	0.27	0.6064	0.49	0.6879	0.71	0.7611	0.93	0.8238
0.06	0.5239	0.28	0.6103	0.50	0.6915	0.72	0.7642	0.94	0.8264
0.07	0.5279	0.29	0.6141	0.51	0.6950	0.73	0.7673	0.95	0.8289
0.08	0.5319	0.30	0.6179	0.52	0.6985	0.74	0.7704	0.96	0.8315
0.09	0.5359	0.31	0.6217	0.53	0.7019	0.75	0.7734	0.97	0.8340
0.10	0.5398	0.32	0.6255	0.54	0.7054	0.76	0.7764	0.98	0.8365
0.11	0.5438	0.33	0.6293	0.55	0.7088	0.77	0.7794	0.99	0.8389
0.12	0.5478	0.34	0.6331	0.56	0.7123	0.78	0.7823	1.00	0.8413
0.13	0.5517	0.35	0.6368	0.57	0.7157	0.79	0.7852	1.01	0.8438
0.14	0.5557	0.36	0.6406	0.58	0.7190	0.80	0.7881	1.02	0.8461
0.15	0.5596	0.37	0.6443	0.59	0.7224	0.81	0.7910	1.03	0.8485
0.16	0.5636	0.38	0.6480	0.60	0.7257	0.82	0.7939	1.04	0.8508
0.17	0.5675	0.39	0.6517	0.61	0.7291	0.83	0.7967	1.05	0.8531
0.18	0.5714	0.40	0.6554	0.62	0.7324	0.84	0.7995	1.06	0.8554
0.19	0.5753	0.41	0.6591	0.63	0.7357	0.85	0.8023	1.07	0.8577
0.20	0.5793	0.42	0.6628	0.64	0.7389	0.86	0.8051	1.08	0.8599
0.21	0.5832	0.43	0.6664	0.65	0.7422	0.87	0.8078	1.09	0.8621

z	F(z)	z	F(z)	z	F(z)	z	F(z)	z	F(z)
1.10	0.8643	1.40	0.9192	1.70	0.9554	2.00	0.9772	2.30	0.9893
1.11	0.8665	1.41	0.9207	1.71	0.9564	2.01	0.9778	2.31	0.9896
1.12	0.8686	1.42	0.9222	1.72	0.9573	2.02	0.9783	2.32	0.9898
1.13	0.8708	1.43	0.9236	1.73	0.9582	2.03	0.9788	2.33	0.9901
1.14	0.8729	1.44	0.9251	1.74	0.9591	2.04	0.9793	2.34	0.9904
1.15	0.8749	1.45	0.9265	1.75	0.9599	2.05	0.9798	2.35	0.9906
1.16	0.8770	1.46	0.9279	1.76	0.9608	2.06	0.9803	2.36	0.9909
1.17	0.8790	1.47	0.9292	1.77	0.9616	2.07	0.9808	2.37	0.9911
1.18	0.8810	1.48	0.9306	1.78	0.9625	2.08	0.9812	2.38	0.9913
1.19	0.8830	1.49	0.9319	1.79	0.9633	2.09	0.9817	2.39	0.9916
1.20	0.8849	1.50	0.9332	1.80	0.9641	2.10	0.9821	2.40	0.9918
1.21	0.8869	1.51	0.9345	1.81	0.9649	2.11	0.9826	2.41	0.9920
1.22	0.8888	1.52	0.9357	1.82	0.9656	2.12	0.9830	2.42	0.9922
1.23	0.8907	1.53	0.9370	1.83	0.9664	2.13	0.9834	2.43	0.9925
1.24	0.8925	1.54	0.9382	1.84	0.9671	2.14	0.9838	2.44	0.9927
1.25	0.8944	1.55	0.9394	1.85	0.9678	2.15	0.9842	2.45	0.9929
1.26	0.8962	1.56	0.9406	1.86	0.9686	2.16	0.9846	2.46	0.9931
1.27	0.8980	1.57	0.9418	1.87	0.9693	2.17	0.9850	2.47	0.9932
1.28	0.8997	1.58	0.9429	1.88	0.9699	2.18	0.9854	2.48	0.9934
1.29	0.9015	1.59	0.9441	1.89	0.9706	2.19	0.9857	2.49	0.9936
1.30	0.9032	1.60	0.9452	1.90	0.9713	2.20	0.9861	2.50	0.9938
1.31	0.9049	1.61	0.9463	1.91	0.9719	2.21	0.9864	2.51	0.9940
1.32	0.9066	1.62	0.9474	1.92	0.9726	2.22	0.9868	2.52	0.9941
1.33	0.9082	1.63	0.9484	1.93	0.9732	2.23	0.9871	2.53	0.9943
1.34	0.9099	1.64	0.9495	1.94	0.9738	2.24	0.9875	2.54	0.9945
1.35	0.9115	1.65	0.9505	1.95	0.9744	2.25	0.9878	2.55	0.9946
1.36	0.9131	1.66	0.9515	1.96	0.9750	2.26	0.9881	2.56	0.9948
1.37	0.9147	1.67	0.9525	1.97	0.9756	2.27	0.9884	2.57	0.9949
1.38	0.9162	1.68	0.9535	1.98	0.9761	2.28	0.9887	2.58	0.9951
1.39	0.9177	1.69	0.9545	1.99	0.9767	2.29	0.9890	2.59	0.9952

z	F(z)	z	F(z)	z	F(z)	z	F(z)	z	F(z)
2.60	0.9953	2.90	0.9981	3.20	0.9993	3.50	0.9998	3.80	0.9999
2.61	0.9955	2.91	0.9982	3.21	0.9993	3.51	0.9998	3.81	0.9999
2.62	0.9956	2.92	0.9982	3.22	0.9994	3.52	0.9998	3.82	0.9999
2.63	0.9957	2.93	0.9983	3.23	0.9994	3.53	0.9998	3.83	0.9999
2.64	0.9959	2.94	0.9984	3.24	0.9994	3.54	0.9998	3.84	0.9999
2.65	0.9960	2.95	0.9984	3.25	0.9994	3.55	0.9998	3.85	0.9999
2.66	0.9961	2.96	0.9985	3.26	0.9994	3.56	0.9998	3.86	0.9999
2.67	0.9962	2.97	0.9985	3.27	0.9995	3.57	0.9998	3.87	0.9999
2.68	0.9963	2.98	0.9986	3.28	0.9995	3.58	0.9998	3.88	0.9999
2.69	0.9964	2.99	0.9986	3.29	0.9995	3.59	0.9998	3.89	0.9999
2.70	0.9965	3.00	0.9987	3.30	0.9995	3.60	0.9998	3.90	1.0000
2.71	0.9966	3.01	0.9987	3.31	0.9995	3.61	0.9998	3.91	1.0000
2.72	0.9967	3.02	0.9987	3.32	0.9995	3.62	0.9999	3.92	1.0000
2.73	0.9968	3.03	0.9988	3.33	0.9996	3.63	0.9999	3.93	1.0000
2.74	0.9969	3.04	0.9988	3.34	0.9996	3.64	0.9999	3.94	1.0000
2.75	0.9970	3.05	0.9989	3.35	0.9996	3.65	0.9999	3.95	1.0000
2.76	0.9971	3.06	0.9989	3.36	0.9996	3.66	0.9999	3.96	1.0000
2.77	0.9972	3.07	0.9989	3.37	0.9996	3.67	0.9999	3.97	1.0000
2.78	0.9973	3.08	0.9990	3.38	0.9996	3.68	0.9999	3.98	1.0000
2.79	0.9974	3.09	0.9990	3.39	0.9997	3.69	0.9999	3.99	1.0000
2.80	0.9974	3.10	0.9990	3.40	0.9997	3.70	0.9999		
2.81	0.9975	3.11	0.9991	3.41	0.9997	3.71	0.9999		
2.82	0.9976	3.12	0.9991	3.42	0.9997	3.72	0.9999		
2.83	0.9977	3.13	0.9991	3.43	0.9997	3.73	0.9999		
2.84	0.9977	3.14	0.9992	3.44	0.9997	3.74	0.9999		
2.85	0.9978	3.15	0.9992	3.45	0.9997	3.75	0.9999		
2.86	0.9979	3.16	0.9992	3.46	0.9997	3.76	0.9999		
2.87	0.9979	3.17	0.9992	3.47	0.9997	3.77	0.9999		
2.88	0.9980	3.18	0.9993	3.48	0.9997	3.78	0.9999		
2.89	0.9981	3.19	0.9993	3.49	0.9998	3.79	0.9999		

Taula 2. Punts de tall de la funció de distribució khi quadrat

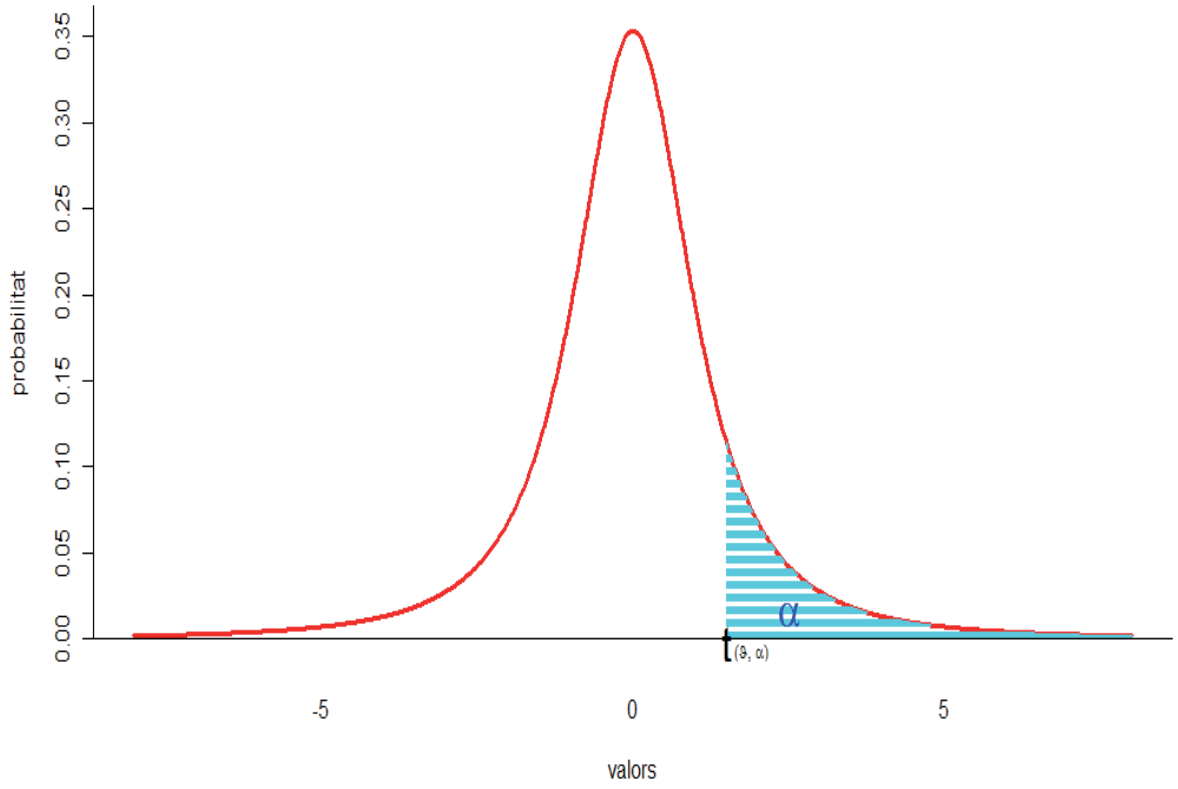


	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	3.9e-6	1.57e-5	9.82e-4	3.9e-3	0.00158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	4.61	5.99	7.38	9.21	10.6
3	0.0717	0.1148	0.2158	0.352	0.584	6.25	7.81	9.35	11.34	12.8
4	0.2070	0.2971	0.4844	0.711	1.064	7.78	9.49	11.14	13.28	14.9
5	0.4117	0.5543	0.8312	1.145	1.610	9.24	11.07	12.83	15.09	16.7
6	0.6757	0.8721	1.2373	1.635	2.204	10.64	12.59	14.45	16.81	18.5
7	0.9893	1.2390	1.6899	2.167	2.833	12.02	14.07	16.01	18.48	20.3
8	1.3444	1.6465	2.1797	2.733	3.490	13.36	15.51	17.53	20.09	22.0

9	1.7349	2.0879	2.7004	3.325	4.168	14.68	16.92	19.02	21.67	23.6
10	2.1559	2.5582	3.2470	3.940	4.865	15.99	18.31	20.48	23.21	25.2
11	2.6032	3.0535	3.8157	4.575	5.578	17.28	19.68	21.92	24.72	26.8
12	3.0738	3.5706	4.4038	5.226	6.304	18.55	21.03	23.34	26.22	28.3
13	3.5650	4.1069	5.0088	5.892	7.042	19.81	22.36	24.74	27.69	29.8
14	4.0747	4.6604	5.6287	6.571	7.790	21.06	23.68	26.12	29.14	31.3
15	4.6009	5.2293	6.2621	7.261	8.547	22.31	25.00	27.49	30.58	32.8
16	5.1422	5.8122	6.9077	7.962	9.312	23.54	26.30	28.85	32.00	34.3
17	5.6972	6.4078	7.5642	8.672	10.085	24.77	27.59	30.19	33.41	35.7
18	6.2648	7.0149	8.2307	9.390	10.865	25.99	28.87	31.53	34.81	37.2
19	6.8440	7.6327	8.9065	10.117	11.651	27.20	30.14	32.85	36.19	38.6
20	7.4338	8.2604	9.5908	10.851	12.443	28.41	31.41	34.17	37.57	40.0
21	8.0337	8.8972	10.2829	11.591	13.240	29.62	32.67	35.48	38.93	41.4
22	8.6427	9.5425	10.9823	12.338	14.041	30.81	33.92	36.78	40.29	42.8
23	9.2604	10.1957	11.6886	13.091	14.848	32.01	35.17	38.08	41.64	44.2
24	9.8862	10.8564	12.4012	13.848	15.659	33.20	36.42	39.36	42.98	45.6
25	10.5197	11.5240	13.1197	14.611	16.473	34.38	37.65	40.65	44.31	46.9
26	11.1602	12.1981	13.8439	15.379	17.292	35.56	38.89	41.92	45.64	48.3
27	11.8076	12.8785	14.5734	16.151	18.114	36.74	40.11	43.19	46.96	49.6
28	12.4613	13.5647	15.3079	16.928	18.939	37.92	41.34	44.46	48.28	51.0
29	13.1211	14.2565	16.0471	17.708	19.768	39.09	42.56	45.72	49.59	52.3
30	13.7867	14.9535	16.7908	18.493	20.599	40.26	43.77	46.98	50.89	53.7
40	20.7065	22.1643	24.4330	26.509	29.051	51.81	55.76	59.34	63.69	66.8
50	27.9907	29.7067	32.3574	34.764	37.689	63.17	67.50	71.42	76.15	79.5
60	35.5345	37.4849	40.4817	43.188	46.459	74.40	79.08	83.30	88.38	92.0
70	43.2752	45.4417	48.7576	51.739	55.329	85.53	90.53	95.02	100.43	104.2
80	51.1719	53.5401	57.1532	60.391	64.278	96.58	101.88	106.63	112.33	116.3
90	59.1963	61.7541	65.6466	69.126	73.291	107.57	113.15	118.14	124.12	128.3
100	67.3276	70.0649	74.2219	77.929	82.358	118.50	124.34	129.56	135.81	140.2

Per exemple, la probabilitat que una variable aleatòria khi quadrat amb 10 graus de llibertat siga superior a 15,99 és de 0,100.

Taula 3. Punts de tall de la distribució *t* de Student

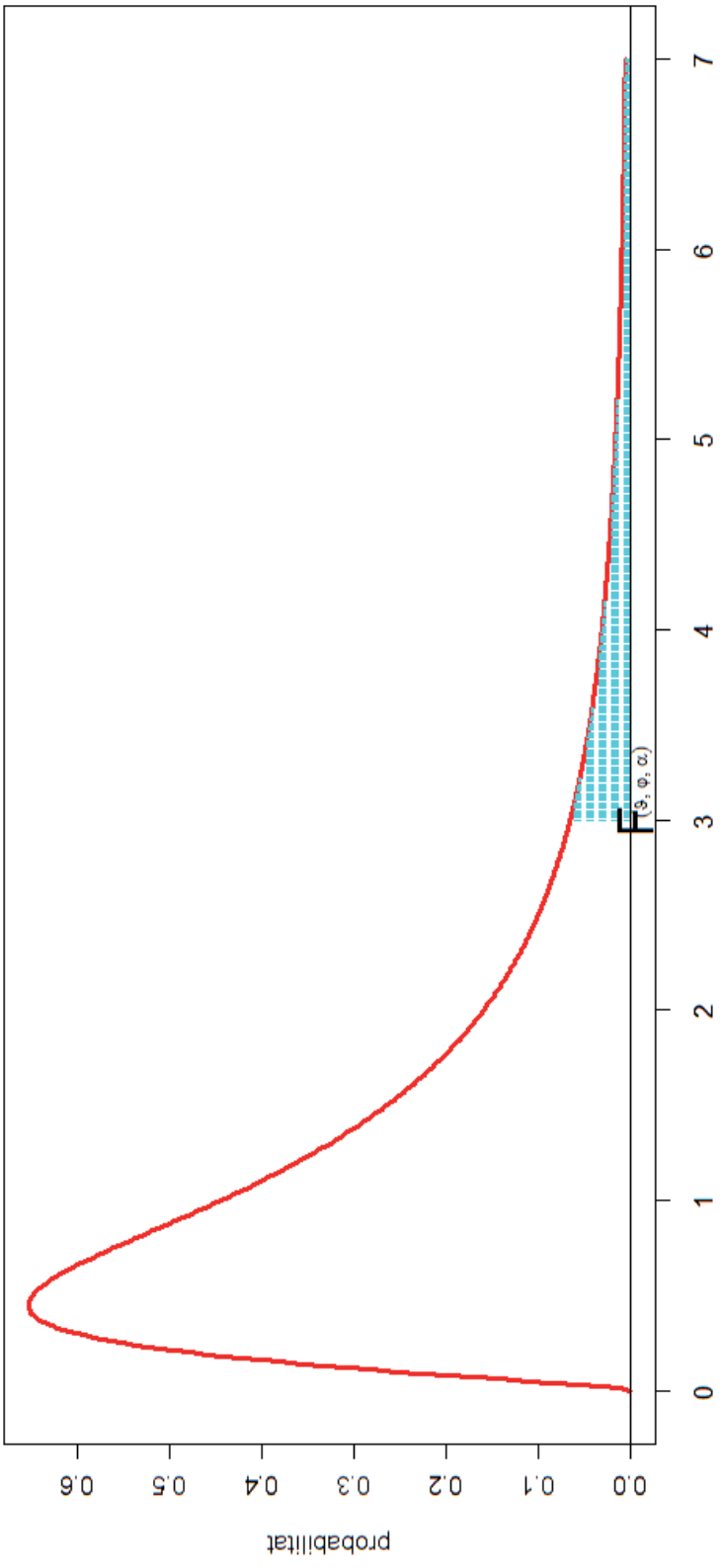


Graus de llibertat (ϑ)	α				
	0.1	**0.05**	**0.025**	**0.01**	**0.005**
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055

13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
31	1.309	1.696	2.040	2.453	2.744
32	1.309	1.694	2.037	2.449	2.738
33	1.308	1.692	2.035	2.445	2.733
34	1.307	1.691	2.032	2.441	2.728
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
45	1.301	1.679	2.014	2.412	2.690
50	1.299	1.676	2.009	2.403	2.678
55	1.297	1.673	2.004	2.396	2.668
60	1.296	1.671	2.000	2.390	2.660
70	1.294	1.667	1.994	2.381	2.648
80	1.292	1.664	1.990	2.374	2.639
90	1.291	1.662	1.987	2.368	2.632
>100	1.290	1.660	1.984	2.364	2.626

Per exemple, la probabilitat que una variable aleatòria *t* de Student amb 10 graus de llibertat siga superior a 1,732 és de 0,10.

Taula 4. Punts de tall de la distribució F de Snedecor



Per exemple, la probabilitat que una variable $F_{3,7}$ siga superior a 4,35 és de 0,05.

$\alpha = 0{,}01$

Denominador (φ)	Numerador (ϑ)																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	
1	4052.2	4999.5	5403.6	5624.6	5763.7	5859.0	5928.7	5981.1	6022.5	6055.9	6106.3	6157.3	6208.7	6234.6	6260.6	6286.7	6313.0	6339.3	6365.9
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00

15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Modesto J. Beltrán/María José Peris - ISBN: 978-84-693-0145-6

$\alpha = 0{,}025$																			
Denominador (φ)	Numerador (ϑ)																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	
1	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
2	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
3	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
4	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
5	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
6	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.41	4.36	4.31	4.25	4.20	4.14
7	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
8	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
9	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
10	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
11	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
12	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
13	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
14	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40

15	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
16	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
17	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
18	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
19	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
20	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
21	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
22	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
23	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
24	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
25	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
26	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
27	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
28	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
29	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
30	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
40	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
60	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
120	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00
	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

$\alpha = 0{,}05$

Denominador (φ)	Numerador (ϑ)																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.0	248.0	249.1	250.1	251.2	252.2	253.2	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13

15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

$\alpha = 0{,}1$

Denominador (φ)	Numerador (ϑ)																		
	1	2	3	4	5	6	7	8	9	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80

Modesto J. Beltrán/María José Peris - ISBN: 978-84-693-0145-6

15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

Bibliografia

Alvira, F. (1989): *Diseños de investigación social: criterios operativos*, en García Ferrando et al. (comps.): *El análisis de la realidad social*, Alianza Universitaria, Madrid.

Barbancho, A. G. (1986): *Estadística elemental moderna*, Ariel Economía, Barcelona.

Biosca, A., M. J. Espinet, M. J. Fandos, M. Jimeno i J. Villagrà (1999): *Matemáticas aplicadas a las Ciencias Sociales II*, Edebé, Barcelona.

Booth, t. i M. Ainscow (2004): *Índice de inclusión. Desarrollando el aprendizaje y la participación en las escuelas***,** Centre for Studies on Inclusive Education, Bristol.

Brunet, I., À. Belzunegui i I. Pastor (2000): *Les tècniques d'investigació social i la seva aplicació*, Universitat Rovira i Virgili, Tarragona.

Bunge, M. (1997): *Las teorías de la causalidad*: Sígueme, Salamanca.

Campbell, D. T. i J. Stanley (1970): *Diseños experimentales y cuasiexperimentales en la investigación social*, Amorrortu, Buenos Aires.

Cea d'Ancona, M. A. (1996): *Metodología cuantitativa. Estrategias y técnicas de investigación social*, Síntesis, Madrid.

Colera, J., R. García i M. J. Oliveira (2003): *Matemàtiques aplicades a les Ciències Socials*, Anaya, Madrid.

Correa, J. C. i N. González (2002): *Gráficos estadísticos con R*, Universidad Nacional de Colombia (sede Medellín), Disponible en: http://cran.r-project.org/doc/contrib/grafi3.pdf.

Instituto de Estudios Políticos (1976): *Diccionario de las ciencias sociales*, 2 volums, Madrid.

Escuder Valles, R. (1987): *Métodos estadísticos aplicados a la economía*, Ariel Economía, Barcelona.

Fernández Cuesta, F. i F. Fuentes García (1994): *Curso de estadística descriptiva. Teoría y práctica*, Ariel, Barcelona.

Filstead, W. J. (1986): *Métodos cualitativos: una experiencia necesaria en la investigación evaluativa*, Morata, Madrid

García Ferrando, M. (1985): *Socioestadística*, Alianza Universidad, Madrid

González Blasco, P. (1989): *Medir en las ciencias sociales*. En García Ferrando, M. et al. (comps.): *El análisis de la realidad social*, Alianza Universitaria, Madrid.

Gracia, F., J. Mateu i P. Vindel (1997): *Problemas de probabilidad y estadística*, Tilde, València.

Hoaglin, D. C. (2003): «John W. Tukey and Data Analysis», *Hoaglin Statistical Science*.

Ibáñez, M. V. i A. Simó (2002): *Apuntes de estadística para Ciencias Empresariales*, Publicacions de la Universitat Jaume I, Castelló de la Plana.

Johnson, R. A. (1997): *Probabilidad y estadística para ingenieros de Miller y Freund*, Prentice-Hall Hispanoamericana, Mèxic.

Kazmier, L. (1998): *Estadística aplicada a la administración y a la economía*, McGraw-Hill, Colombia, 3a edició.

Loève, M. (1976): *Teoría de la probabilidad*, Tecnos, Madrid.
Mark Sirkin, R. (2006): *Statistics for the Social Sciences*, Sage, Thousand Oak, Califòrnia, 3a edició.
Martín Pliego, J. (1995): *Introducción a la estadística económica y empresarial*, Editorial ac, Madrid.
Martín, P. i J. Martín Pliego (1991): *Curso básico de estadística económica*, Editorial ac, Madrid, 3a edició.
Meyer, P. L. (1986): *Probabilidad y aplicaciones estadísticas*, Addison-Wesley, eua.
Monteagudo, M. F. i J. Paz (2003): *Matemáticas aplicadas a las Ciencias Sociales II*, Luis Vives, Saragossa.
Montero Lorenzo, J. M. (2003): *Estadística para relaciones laborales*, Editorial AC, Madrid.
Moore, D. S., G. P. McCabe, W. M. Duckworth i S. L. Sclove (2003): *The Practice of Business Statistics, Using Data for Decisions*, W. H. Freeman and Company, eua.
Newbold, P., W. L. Carlson i B. Thorne (2007): *Estadística para administración y economía*, Prentice Hall, Madrid.
Polit San Román, A. (1981): *Estadística i tècniques d'investigació* social, Piràmide, Madrid.
Ritzer, G. (1993): *Teoría de la sociología contemporánea,* MC Graw Hill, Madrid.
Rodríguez Ibáñez, E. (1989): *La perspectiva sociológica. Historia, teoría y método*, Taurus, Madrid.
Ruiz-Maya, L. i F. J. Martín-Pliego (2005): *Fundamentos de inferencia estadística*, Thompson, Madrid, 3a edició.
Sanz, J. A., A. Bedate, A. Vivas i J. González (1996): *Problemas de estadística descriptiva empresarial*, Ariel Economía, Barcelona.
Sierra Bravo, R. (1995): *Técnicas de investigación social. Teoría y ejercicios*, Paraninfo, Madrid.
Singh, S. (1997): *Fermat's Enigma*, Anchor, Nova York.
Spiegel, Murray R. (1991): *Estadística*, McGraw-Hill, Madrid.
Schwartz, H. i J. Jacobs (1984): *Sociología cualitativa: método para la construcción de la realidad*, Trilles, Mèxic.
Tomeo Perucha, V. i I. Uña Juárez (2003): *Diez lecciones de estadística descriptiva (curso teórico-práctico)*, Editorial ac, Madrid.
Triola, M. F. (2000): *Estadística elemental*, Pearson Education, Mèxic, 7a edició.
Uña Juárez, I., V. Tomeu Perucha i J. San Martín Moreno (2003): *Lecciones de cálculo de probabilidad, curso teórico-práctico*, Thompson, Madrid.
Venables, W. N., D. M. Smith i The R Development Core Team (2007): *An Introduction to R*, R Foundation for Statistical Computing, Viena.
Webster, A. L. (2000): *Estadística aplicada a los negocios y a la economía*, McGraw-Hill, Colòmbia.
Wonnacot, T. H. i R. J. Wonnacot (1996): *Introducción a la estadística*, Limusa Noriega Editores, Mèxic.
Zaiats, V., M. L. Calle i R. Presas (1998): *Probabilitat i estadística. Exercicis I*, Eumo, Barcelona.